THESE de DOCTORAT D'ETAT

ès Sciences Mathématiques

présentée

à l'Université Pierre et Marie Curie
- Paris VI -

par Monsieur Pierre MICHAUD
pour obtenir le grade de DOCTEUR ès SCIENCES
Sujet de la thèse :

AGREGATION DE PREFERENCES

soutenue le 25 Juin 1981
devant le jury composé de :

Messieurs	D.	DUGUE	*(président)*
	P.	DEHEUVELS	*(rapporteur)*
	J.	GEFFROY	*(rapporteur)*
	G.	KREWERAS	*(rapporteur)*
	M.	BARBUT	*(examinateur)*
	I. C.	LERMAN	*(invité)*

UNIV.
LEIDEN
BIBL.

REMERCIEMENTS

Cette thèse présentée à l'Université Paris VI a été réalisée sous la direction du Professeur D. DUGUE, Directeur de l'I.S.U.P.. Je remercie le Professeur D. DUGUE de l'honneur qu'il me fait en présidant le jury de ma thèse.

Je remercie très vivement le Professeur J.GEFFROY, Directeur du 3ème cycle "Statistiques et Recherche Opérationnelle" de l'Université Paris VI, d'avoir accepté la charge de rapporteur de la thèse et de m'avoir, à cette occasion donné la possibilité d'enseigner dans le cadre de son D.E.A..

Je suis d'autre part extrêmement reconnaissant au Professeur G. KREWERAS de l'Université Paris II, d'avoir accepté d'être, rapporteur de la thèse que je présente. Ses suggestions constructives m'ont été fort utiles.

Je remercie le Professeur P. DEHEUVELS de l'Université Paris VI qui, ayant assisté à la progression de ces travaux et s'y étant favorablement interessé, a bien voulu accepter d'être également rapporteur.

Toute ma gratitude va aussi au Professeur M. BARBUT, Vice-Président de l'Université Paris V, l'un des premiers en France à s'être interessé à ce type de problème, pour sa participation au jury de cette thèse.

Mes remerciements vont également au Professeur I.C. LERMAN de l'Université de Rennes II d'avoir accepté l'invitation qui lui a été faite d'assister à la soutenance de cette thèse.

Par ailleurs, je tiens à profiter de l'occasion qui m'est donnée ici pour remercier à nouveau très chaleureusement Monsieur R. MOREAU, Directeur du Développement Scientifique IBM-France, d'avoir, depuis mon arrivée au sein de son Service, suivi avec le plus grand interêt le déroulement de mes travaux, de m'avoir incité à publier et judicieusement conseillé dans le choix d'applications de ces méthodes.

Je suis également très redevable à Monsieur M. PAPO, Directeur Scientifique de la Compagnie IBM-France, de la bienveillance qu'il a constamment manifestée et ce, dès l'origine, à l'égard de mes recherches en Analyse des Données et d'en avoir favorisé la réalisation tant sur le plan matériel que scientifique.

Je me dois d'associer à ces remerciements les autres

personnalités scientifiques qui, les premières, se sont interessées à ces travaux. Je citerai ici plus particulièrement A.J. HOFFMAN du Laboratoire de Recherche IBM de Yorktown Heights, ainsi que B. MONJARDET de l'Ecole des Hautes Etudes en Sciences Sociales.

Pour terminer, mes remerciements vont à toutes les personnes qui, d'une façon ou d'une autre, ont permis la concrétisation de cette thèse; je pense ici en parliculier aux spécialistes du système MPSX/370 M. BENICHOU, J.M. GAUTHIER, G. RIBIERE et bien entendu à mon collègue et ami F. MARCOTORCHINO.

PRESENTATION DE LA THESE

Cette thèse présente un certain nombre de résultats concernant le problème d'agrégation d'opinions.
Les opinions considérées sont en général des opinions de préférences, le cas d'opinions de similarités étant traité dans la thèse de J.F. Marcotorchino [6].

Le type de problème d'agrégation d'opinions que nous condidérons est le suivant:
m "juges" donnent chacun leur opinion sur n "objets". Il s'agit de définir une opinion "collective", sur les n objets, à partir des m opinions "individuelles" à l'aide d'une règle d'agrégation. Intuitivement, l'opinion collective doit représenter "le mieux possible" l'ensemble des opinions individuelles.
Le cas le plus connu est certainement celui où l'opinion de chaque juge correspond à des préférences exprimées sous la forme d'un classement des n objets (ordre total). C'est le problème de l'agrégation de classements.
Il s'agit alors de représenter l'ensemble des m classements des différents juges par un classement collectif, à l'aide d'une règle d'agrégation.

L'exemple le plus célèbre qui est à l'origine de ce type de problème date du XVIIIème siècle. Il s'agit du problème "d'élections" où chaque "juge" représente un votant et chaque "objet" un candidat à élire; le classement collectif représentant alors l'ordre d'élection des candidats. Ce problème a été posé en 1781 par Borda qui, à cette occasion, proposa la règle d'agrégation désormais classique qui porte son nom [2].

Dès l'origine, différentes difficultés inhérentes à ce type de problèmes sont apparues:

- d'une part, l'existence de différentes règles d'agrégation pour traiter ce problème.
Mais ces différentes règles d'agrégation donnent, en général, des résultats (opinion collective) différents. Dans ces conditions, on peut se demander si tous les résultats obtenus sont valables ou bien s'il y a des règles meilleures que d'autres.

- d'autre part, l'existence de paradoxes.
Dans ce cas, l'existence même d'une seule règle d'agrégation satisfaisante devient problématique.
Par exemple, la règle d'agrégation de Borda a été critiquée par Condorcet. En 1785, Condorcet proposa un principe

d'agrégation différent de celui de Borda, reposant sur la règle de la majorité et le principe de comparaisons par paires.
Malheureusement, comme l'a remarqué Condorcet lui-même, les opinions individuelles peuvent contenir un "effet Condorcet" et dans ce cas la "méthode d'agrégation" ne permet pas de déduire un classement collectif.

Depuis Borda et Condorcet, on sait que l'agrégation des préférences est un problème délicat. Et depuis cette époque, la définition d'une "bonne" règle d'agrégation est un des thèmes principaux des chercheurs tant français qu'étrangers.

Mais la définition d'une règle d'agrégation peut se faire de différentes façons. Très schématiquement, on pourra considérer deux tendances principales:

1) les définitions "constructives". Dans ce cas, la définition de la règle permet d'obtenir une opinion collective en un nombre fini de calculs élémentaires;

2) les définitions "descriptives" ou "axiomatiques". Il s'agit d'une définition en quelque sorte indirecte; la règle est "décrite" par la définition de conditions que devra vérifier la règle en question. Mais ces conditions ne définissent pas nécessairement une règle unique, et selon les cas on pourra avoir:
- aucune règle (condition d'impossibilité),
- une seule règle (caractérisation axiomatique),
- plusieurs règles.

Depuis quelques dizaines d'années cette dernière approche, à la suite des travaux d'Arrow, a la faveur de bon nombre de chercheurs. Bien entendu, dans une telle perspective le calcul effectif d'une opinion collec ive, pour une règle donnée, devient secondaire.

Dans la première approche, "obtention d'une opinion collective en un nombre fini de calculs élémentaires" ne veut pas nécessairement dire "opinion collective effectivement calculable"; et un grand nombre de règles sont de nature combinatoire. Par exemple, dans le cas d' agrégation de classements de nombreuses règles nécessitent l'évaluation d'un critère pour les différents classements possibles.
Pour 68 objets à classer le nombre de classements à considérer est supérieur à 10^{100}, ce qui rend bien entendu tout calcul explicite absolument impossible même avec un ordinateur extrêmement puissant.

Aussi, existe-t-il une grande différence entre le nombre important de travaux sur le sujet et les règles d'agrégation utilisées pratiquement. Bien entendu, dans la pratique, la possibilité de calcul effectif d'une opinion collective est

essentielle. Bon nombre de règles sont utilisées bien plus en raison de leur simplicité que de leur qualité; alors que d'autres règles beaucoup plus satisfaisantes sont inutilisées à cause de leur complexité. C'est le cas, par exemple, pour les problèmes d'agrégation de tableaux de comparaisons où la règle de la majorité sous contraintes (décrite dans cette thèse), amplement étudiée et considérée comme "très satisfaisante", est pratiquement inutilisée à cause de sa complexité.

Tel est le point de départ de cette thèse qui est axée sur le calcul effectif d'opinions collectives pour diverses règles d'agrégation.

Le principe général en est le suivant:

- chaque règle étudiée est représentée par un modèle d'optimisation;
- pour chaque modèle d'optimisation, la résolution est obtenue par une technique d'optimisation appropriée.

Nous avons considéré 3 grands types de problèmes d'agrégation d'opinion:

1) Agrégation "d'échelles",
3) Agrégation de classements donnés sous forme de rangs,
3) Agrégation de tableaux de comparaisons par paires.

Dans le dernier cas, le terme agrégation est pris au sens large et les tableaux de comparaisons individuels et collectifs peuvent être de nature différente. Les tableaux individuels peuvent être "intransitifs" et le tableau collectif "transitif", comme par exemple dans les tournois sportifs.

En général, une règle est obtenue en définissant une distance entre deux opinions puis un éloignement entre l'opinion collective et l'ensemble des opinions individuelles. On cherchera une opinion collective à éloignement minimal.

Ces trois cas correspondent chacun à des modèles d'optimisation différents:

- dans le premier cas, il s'agit de programmes non linéaires sans contraintes avec solution explicite ou implicite selon le cas;
- dans le deuxième cas, il s'agit d'un programme d'affectation linéaire qui est un programme linéaire sur un réseau d'un type particulier;
- dans le troisième cas, il s'agit d'un programme linéaire à variables bivalentes.

Il est important de noter que dans les cas 2 et 3 il s'agit

en général de problèmes combinatoires pratiquement insolubles et que les méthodes proposées permettent la résolution pratique de ces problèmes.
Ces résultats ont été présentés dans les publications suivantes:

- G. AYACHE, F. MARCOTORCHINO, P. MICHAUD, "Démocratie Française et la presse. Techniques d'analyse de la Réception d'un message", Revue Française de Communication, No 1, 30, (1978).

- F. MARCOTORCHINO, P. MICHAUD, "Méthodes d'Analyse Ordinale des Données", in Actes du Colloque, II, Le Diagnostic d'Entreprise, Publication du C.E.R.M.E., I.A.E. Nice, (1978).

- P. MICHAUD, F. MARCOTORCHINO, [P1] "Optimization in Ordinal Data Analysis, Etude N001, Centre Scientifique I.B.M. France, version augmentée, (1978).

- F. MARCOTORCHINO, P. MICHAUD, [L] "Optimisation en Analyse Ordinale des Données", Masson, Paris, (1979).

- P. MICHAUD, F. MARCOTORCHINO,[P2]"Modèles d'optimisation en Analyse des Données Relationnelles, Mathématiques et Sciences Humaines, No 67, pp. 7 à 38, Dunod, Paris, (1979).

- P. MICHAUD, F. MARCOTORCHINO, "Optimisation en Analyse des Données Relationnelles", in Data Analysis and Informatics, E. Diday et al. (eds), North Holland Publishing Company, (1980).

- F. MARCOTORCHINO, P. MICHAUD, "Optimization in Exploratory Data Analysis", "proceedings" du 5ème Symposium International de R.O., Cologne, Physika Verlag, (1981).

- F. MARCOTORCHINO, P. MICHAUD, "Agrégation des Similarités en Classification Automatique", Etude F012, Centre Scientifique I.B.M. France, (1981). à paraître sous forme simplifiée dans la Revue de Statistiques Appliquées.

- P. MICHAUD, F. MARCOTORCHINO, [P3] "Optimization in Opinion's Aggregation", Etude F016, Centre Scientifique I.B.M. France, (1981). à paraître dans Mathematical Programming.

En ce qui concerne le problème d'agrégation des préférences,

qui est l'objet de cette thèse, on trouvera les principaux résultats dans le livre [L] "Optimisation en Analyse Optimale des Données" publié chez Masson en 1979; livre auquel on devra ajouter les publications suivantes (incluses dans cette thèse):
- [P1]
- [P2]
- [P3]

Le livre [L] étudie les trois types d'agrégation précédemment mentionnés. Dans [P1], l'agrégation de tableaux de comparaisons par paires est généralisée au cas de préordres totaux et de partitions. [P2] considère des cas encore plus généraux comme l'agrégation pour un semi-ordre. Enfin, [P3] est un article de synthèse destiné au public anglo-saxon.

Parmi les principaux résultats obtenus personnellement j'indiquerai:
pour la partie agrégation de classements donnés sous forme de rangs, traités en [L] et [P1], une modélisation générale sous forme linéaire de l'ensemble des règles étudiées, ainsi que de nouvelles méthodes de résolution. En contrepartie dans ce domaine, F. Marcotorchino s'est surtout intéressé à la définition d'indicateurs de concordance et à leur utilisation pour définir des problèmes d'agrégation. Comme l'indique le titre du livre [L] "Optimisation en Analyse Ordinale des Données", les problèmes traités ne sont pas strictement limités au cadre de l'agrégation et l'on trouvera aussi le problème en quelque sorte inverse, où la règle et l'opinion collective étant fixée, on recherche les opinions individuelles. Il s'agit d'une "régression ordinale". Là encore, les résultats sont obtenus par programmation linéaire.

En ce qui concerne la partie agrégation de tableaux de comparaisons par paires, je me suis à nouveau intéressé aux méthodes de résolution par programmation linéaire. En ce qui concerne la modélisation linéaire bivalente, partant de l'agrégation de classements, traitée dans [L], nous avons conjointement, F. Marcotorchino et moi-même, étendu dans [P1] cette modélisation aux cas de préordres totaux et de partitions.
Dans [P2] j'ai donné d'autres modélisations linéaires pour d'autres types de problèmes, en particulier l'agrégation par un semi-ordre. Toujours dans [P2], j'ai également obtenu de nouvelles "conditions paretiennes" pour des problèmes d'agrégation où les tableaux individuels et collectifs sont de nature différente, cas très fréquent dans la pratique.
J'ai également proposé dans [L] une méthode heuristique pour le problème d'agrégation de tableaux de comparaisons par paires par un ordre total. Parallèlement, F. Marcotorchino s'est surtout intéressé aux structures et propriétés algébriques de ce même problème.
Ces différentes modélisations m'ont également amené à

considérer des équivalences de modèles ou de règles, comme dans [L] ou [P1], dans le but de simplifier les calculs. Mais l'on peut aussi s'intéresser à des équivalences de règles du point de vue de l'interprétation pratique. C'est ce que j'ai fait avec des "règles régularisées".
Comme je l'ai déjà mentionné F. Marcotorchino étudie dans [L] et [P1], ainsi que dans sa thèse [6], le problème de la concordance d'opinions. J'ai également abordé ce problème et étudié son lien avec l'agrégation d'opinions, mais dans une perspective différente de celle de F. Marcotorchino. A cette occasion, j'ai proposé des indicateurs de cohérence mesurant la qualité d'une agrégation.

Dans cette présentation j'ai volontairement limité les références au strict minimum, chaque publication mentionnée contenant ses propres références. Une liste des auteurs ayant également travaillé sur le sujet m'a parue superflue dans la mesure où elle existait déjà ailleurs. On en trouvera une en particulier dans [L] ou bien dans le No 67 de la Revue Mathématiques et Sciences Humaines contenant l'article [P2]; cette dernière liste étant proposée par B.Monjardet [7].

Pour terminer j'indiquerai brièvement quelques résultats qui ne se trouvent ni dans [L] ni dans [P1], [P2] ou [P3] mais apparaissent sous forme dispersée dans les différentes autres publications.

Il s'agit des résultats suivants:

A) Temps de calcul avec les méthodes heuristiques,
B) Règles régularisées,
C) Indicateurs de cohérence.

A) TEMPS DE CALCUL

Nous donnons ici quelques temps de résolution obtenus avec un ordinateur IBM 168/370; le temps de calcul comprenant les phases de génération, résolution et édition.

Dans la colonne NATURE "P" signifie que le résultat obtenu est une "partition", "C" que le résultat obtenu est un "classement".

Les temps de la colonne méthode exacte ont été obtenus en utilisant le système de programmation mathématique IBM MPSX/370.

Les temps de la colonne heuristique correspondent à l'heuristique en agrégation des préférences pour les exemples "C" (voir [L]) et à l'heuristique en agrégation des

similarités pour les exemples "P" (voir thèse F.Marcotorchino[6]).

TYPE du PB	TAILLE	NATURE	TEMPS(Meth.Exact.)	TEMPS(heuri.)
Archeologie	21	P	7 s	0.1 s
Sociologie	20	P	13 s	0.1 s
Medecine	23	P	15 s	0.2 s
Linguistique	25	C	7 s	0.2 s
Linguistique	28	C	11 s	0.2 s
Medecine	29	P	22 s	0.2 s
Linguistique	29	C	12 s	0.2 s
Esthetique	36	C	22 s	0.2 s
Psychologie	35	P	36 s	0.25 s
Sociologie	40	C	31 s	0.2 s
Test Compar.	40	P	297 s	0.35 s
Linguistique	45	C	60 s	0.25 s
Linguistique	45	C	58 s	0.25 s
Esthétique	48	C	65 s	0.30 s
Esthétique	54	C	117 s	0.30 s
Politique	54	P	220 s	0.90 s
Sociologie	72	C	157 s	0.90 s
Linguistique	77	C	331 s	1.20 s

B) FONCTIONS ECONOMIQUES REGULARISEES

Rappelons, tout d'abord, dans le cas d'agrégation de tableaux de comparaisons quelques définitions.
Le tableau de comparaisons par paires C^k d'un juge k est défini par:

c^k_{ij}=1 si i est en relation avec j pour le juge k,
=0 sinon.

Le tableau de comparaisons C est défini par:

c_{ij}=nombre de juges ayant mis i en relation avec j.

Le tableau de comparaisons C' est défini par:

c'_{ij}=nombre de juges ayant mis i en relation avec j moins nombre de juges n'ayant pas mis i en relation avec j.

Cette définition de C' est valable même dans le cas de comparaisons manquantes.
Le tableau de comparaisons collectif Y est défini par:

y_{ij}=1 si i est en relation avec j dans le tableau collectif,
=0 sinon.

Dans tous ces tableaux on ne compare jamais i avec lui-même.
Le critère d'agrégation à maximiser est défini par:

$$\sum_{i \neq j} c'_{ij} y_{ij}.$$

Dans le cas où le tableau collectif Y représente un ordre total, la fonction économique

$$\sum_{i \neq j} c'_{ij} \; y_{ij}$$

peut être remplacée par d'autres fonctions économiques équivalentes; c'est à dire: la maximisation pour Y ordre total de cette nouvelle fonction économique admet les mêmes solutions optimales que le problème initial.
Par exemple,

$$\sum_{i \neq j} c_{ij} \; y_{ij}$$

$$-\sum_{i \neq j} c_{ji} \; y_{ij}$$

et

$$\sum_{i \neq j} (c_{ij} - c_{ji}) \; y_{ij}$$

sont des fonctions économiques équivalentes à:

$$\sum_{i \neq j} c'_{ij} \; y_{ij}$$

lorsque Y est un ordre total. La démonstration en est donnée en [L] et [P1].
J'ai proposé dans [P1] d'autres cas d'équivalence, tant dans le cas où Y est un ordre total que dans celui où Y représente une partition. J'ai également proposé d'autres cas d'équivalence de fonctions économiques. Il s'agit de fonctions économiques régularisées. Ces fonctions régularisées possèdent certaines propriétés très utiles dans les applications pratiques.

1) Fonction économique régularisée dans le cas où Y est un ordre total

Il s'agit de la fonction économique

$$\sum_{i \neq j} c''_{ij} \; y_{ij}$$

où c''_{ij} est défini par:

$$c''_{ij} = \sum_{k} c''^{k}_{ij}$$

avec:

$$c''^{k}_{ij} = 1/2 \; (c'^{k}_{ij} - c'^{k}_{ji})$$

c'est à dire:

$$c''_{ij} = 1/2 \; (c'_{ij} - c'_{ji})$$

Comparons les tableaux C' et C" après regroupement des

cases (i,j) et (j,i).
Quatre situations de comparaisons sont alors possibles plus une, correspondant au cas de "non-réponse", c'est à dire absence de renseignement concernant la comparaison entre i et j.
Pour le tableau C'^k on obtient:

1) équivalence entre i et j $\quad c'^k_{ij} = 1$ et $c'^k_{ji} = 1$

2) incomparabilité entre i et j $\quad c'^k_{ij} = -1$ et $c'^k_{ji} = -1$

3) i préféré strictement à j $\quad c'^k_{ij} = 1$ et $c'^k_{ji} = 1$

4) j préféré strictement à i $\quad c'^k_{ij} = -1$ et $c'^k_{ji} = 1$

5) "non réponse" $\quad c'^k_{ij} = 0$ et $c'^k_{ji} = 0$

Pour le tableau C''^k les mêmes situations sont représentées par:

1) équivalence entre i et j $\quad c''^k_{ij} = 0$ et $c''^k_{ji} = 0$

2) incomparabilité entre i et j $\quad c''^k_{ij} = 0$ et $c''^k_{ji} = 0$

3) i préféré strictement à j $\quad c''^k_{ij} = +1$ et $c''^k_{ji} = -1$

4) j préféré strictement à i $\quad c''^k_{ij} = -1$ et $c''^k_{ji} = +1$

5) "non réponse" $\quad c''^k_{ij} = 0$ et $c''^k_{ji} = 0$

On constate que dans le tableau C''^k une équivalence ou une incomparabilité entre i et j est assimilée à une "non réponse" et représentée par:

$$c''^k_{ij} = 0 \quad \text{et} \quad c''^k_{ji} = 0.$$

Il en résulte que:
c''_{ij} = nombre de juges ayant strictement préféré i à j moins le nombre de juges ayant strictement préféré j à i.
Bien entendu,

$$c''^k_{ij} = c'^k_{ij}$$

dès que l'on a antisymétrie et totalité, c'est à dire "préférence stricte" entre j et i puisque dans ce cas:

$$c'^k_{ij} = -c'^k_{ji} \quad \text{et} \quad c''^k_{ij} = 1/2(2c'^k_{ij})$$

Démonstration de cette équivalence

$$\text{Max } \sum_{i \neq j} c'_{ij} \, Y_{ij}$$

$$y_{ij}+y_{ji}=1$$

$$y_{ij}+y_{jk}-y_{ik} \leq 1 \quad i\neq j \ , j\neq k \ , i\neq k$$

$$y_{ij} = 0 \text{ ou } 1$$

se réduit à:

$$\text{Max} \sum_{i\neq j} c'_{ij}\, y_{ij}$$

$$y_{ij} + y_{ji} = 1 \quad i<j$$

$$0 \leq y_{ij} + y_{jk} - y_{ik} \leq 1 \quad i<j<k$$

$$y_{ij} = 0 \text{ ou } 1 \quad i\neq j$$

On peut alors éliminer les variables y_{ji} et les relations

$$y_{ij} + y_{ji} = 1 \quad i<j$$

on obtient:

$$\text{Max} \sum_{ij>i} \sum (c'_{ij} - c'_{ji})\, y_{ij} + \sum_{ij>i} \sum c'_{ji}$$

$$0 \leq y_{ij} + y_{jk} - y_{ik} \leq 1 \quad i<j<k$$

$$y_{ij} = 0 \text{ ou } 1 \quad i<j$$

Considérons maintenant:

$$\text{Max} \sum_{i\neq j} c''_{ij}\, y_{ij} \quad \text{pour Y ordre total}$$

En procédant comme précédemment, on obtient:

$$\text{Max} \sum_{ij>i} \sum (c''_{ij} - c''_{ji})\, y_{ij} + \sum_{ij>i} \sum c''_{ji}$$

$$0 \leq y_{ij} + y_{jk} - y_{ik} \leq 1 \quad i<j<k$$

$$y_{ij} = 0 \text{ ou } 1 \quad i<j$$

On a bien le résultat annoncé, en remarquant que:

$$c''_{ij} - c''_{ji} = 1/2(c'_{ij} - c'_{ji}) - 1/2(c'_{ji} - c'_{ij}) = c'_{ij} - c'_{ji}$$

2) Fonction économique régularisée dans le cas où Y est une partition

Il s'agit de la fonction économique

$$\sum_{i\neq j} c''_{ij}\, y_{ij}$$

où c''_{ij} est défini par:

$$c''_{ij} = \sum_k c''^{k}_{ij}$$

avec

$$c''^{k}_{ij} = +1/2\,(c'^{k}_{ij} + c'^{k}_{ji})$$

c'est à dire:

$$c''_{ij} = 1/2\,(c'_{ij} + c'_{ji})$$

Comme précédemment, l'étude des différentes situations ou comparaisons possibles pour les tableaux C'^{k} et C''^{k} donne:

- Pour C'^{k} :

1) ressemblance stricte entre i et j
$c'^{k}_{ij} = 1$ et $c'^{k}_{ji} = 1$

2) dissemblance stricte entre i et j
$c'^{k}_{ij} = -1$ et $c'^{k}_{ji} = -1$

3) "similarité asymétrique"
$c'^{k}_{ij} = +1$ et $c'^{k}_{ji} = -1$

4) "similarité asymétrique"
$c'^{k}_{ij} = -1$ et $c'^{k}_{ji} = +1$

5) "non réponse"
$c'^{k}_{ij} = 0$ et $c'^{k}_{ji} = 0$

- Pour C''^{k} :

1) ressemblance stricte entre i et j
$c''^{k}_{ij} = 1$ et $c''^{k}_{ji} = 1$

2) dissemblance stricte entre i et j
$c''^{k}_{ij} = -1$ et $c''^{k}_{ji} = -1$

3) "similarité asymétrique"
$c''^{k}_{ij} = 0$ et $c''^{k}_{ji} = 0$

4) "similarité asymétrique"
$c''^{k}_{ij} = 0$ et $c''^{k}_{ji} = 0$

5) "non réponse"
$c''^{k}_{ij} = 0$ et $c''^{k}_{ji} = 0$

On constate cette fois que, dans le tableau C''^{k}, une "similarité asymétrique" entre i et j est assimilée à une non réponse et est représentée par:

$$c''^{k}_{ij} = 0 \quad \text{et} \quad c''^{k}_{ji} = 0.$$

Il en résulte que:

c''_{ij} = Nombre de juges pour lesquels i et j sont strictement semblables moins le nombre de juges pour lesquels i et j sont strictement dissemblables.
On a aussi

$$c''^{k}_{ij} = c'^{k}_{ij}$$

dès qu'il y a symétrie entre la comparaison i,j et j,i c'est à dire:

$$c'^{k}_{ij} = c'^{k}_{ji}$$

Démonstration de cette équivalence

$$\text{Max} \sum_{i \neq j} c'_{ij} \; y_{ij}$$

$$y_{ij} - y_{ji} = 0 \qquad i \neq j$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \qquad i \neq j \quad j \neq k \quad i \neq k$$

$$y_{ij} = 0 \text{ ou } 1$$

peut s'écrire:

$$\text{Max} \sum_{i<j} c'_{ij} \; y_{ij}$$

$$y_{ij} - y_{ji} = 0 \qquad i<j$$

$$-y_{ij} + y_{jk} + y_{ik} \leq 1$$

$$y_{ij} - y_{jk} + y_{ik} \leq 1 \qquad i<j<k$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1$$

$$y_{ij} = 0 \text{ ou } 1 \qquad i \neq j$$

On peut éliminer les variables y_{ji} pour $i<j$ et les relations

$$y_{ij} - y_{ji} = 0 \qquad i<j$$

on obtient:

$$\text{Max} \sum_{i<j} (c'_{ij} + c'_{ji}) y_{ij}$$

$$-y_{ij} + y_{jk} + y_{ik} \leq 1$$

$$y_{ij} - y_{jk} + y_{ik} \leq 1 \qquad i<j<k \qquad (3)$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1$$

$y_{ij} = 0$ ou $1 \quad i \neq j$

En opérant de même pour

$\text{Max} \sum_{i \neq j} c'_{ij} \; y_{ij}$ pour Y partition

on obtient:

$\text{Max} \sum_{i \neq j} (c''_{ij} + c''_{ji}) y_{ij}$ avec les contraintes (3)

et l'on a bien le résultat annoncé puisque:

$c''_{ij} + c''_{ji} = 1/2(c'_{ij} + c'_{ji}) + 1/2(c'_{ji} + c'_{ij}) = c'_{ij} + c'_{ji}$

Les deux fonctions sont égales.

C) CORRELATION, CONCORDANCE, ACCORD ET COHERENCE

On peut définir le "degré de liaison" entre deux classements à l'aide d'un coefficient de corrélation. C'est un coefficient qui varie de -1 à +1; -1 indiquant deux ordres dont l'un est l'inverse de l'autre, +1 deux ordres identiques.
Deux coefficients de corrélation de rangs classiques sont le coefficient rho de C. Spearman et le coefficient tau de M.G. Kendall.
On peut généraliser cette notion au cas de m classements. C'est ce qu'a fait Kendall [5]. Dans ce cas au coefficient rho correspond le coefficient K que Kendall appelle "coefficient de concordance", tandis qu'au coefficient tau correspond le coefficient U que Kendall appelle "coefficient d'agrément". Notons que ce coefficient U de Kendall donné dans [5] est général. Il n'est pas limité au cas de classements mais valable pour des tableaux de comparaisons par paires quelconques.
Bien entendu d'autres "coefficients de concordance" sont possibles, on en trouvera par exemple dans [L] donnés par F. Marcotorchino.
Mais, contrairement aux coefficients de corrélation, les coefficients de concordance ne varient pas de -1 à +1, mais en principe de 0 à 1. Kendall s'explique sur ce point dans [5] page 95. C'est que, dans le cas de plus de deux classements (ou de deux tableaux de comparaisons par paires), les notions d'accord ou de désaccord ne sont plus symétriques. Par exemple, lors de la comparaison d'un objet i avec un objet j les m juges peuvent être en accord complet, mais un désaccord complet est impossible ainsi, si le juge 1 est en désaccord avec le juge 2 et le juge 2 en désaccord avec le juge 3, il y aura nécessairement accord entre le juge 1 et le juge 3.
Mais, contre toute attente, le coefficient U de Kendall ne varie pas de 0 à 1. Le maximum de ce coefficient est bien égal à 1 lorsqu'il y a unanimité, mais la valeur du minimun est négative. Elle est égale à:

$$\frac{-1}{m} \qquad \text{lorsque m est impair,}$$

et à:

$$\frac{-1}{m-1} \qquad \text{lorsque m est pair}$$

avec la valeur -1 lorsque m = 2.
On aurait une situation analogue, le minimun pouvant être différent de 0, en utilisant des variantes de ce coefficient, voir par exemple la thèse de F. Marcotorchino

Ce résultat n'est guère satisfaisant et il semble souhaitable d'avoir un coefficient variant de 0 à 1 C'est ce que nous proposons maintenant dans le cadre général de m tableaux de comparaisons par paires.

Rappel ons que le tableau de comparaisons par paires C^k d'un juge k est défini par:

c_{ij}^k =1 si i est en relation avec j pour le juge k;
=0 sinon.

Le tableau de comparaison C est défini par:

c_{ij} =nombre de juges ayant mis i en relation avec j.

$m-c_{ij}$ représente le nombre de juges n'ayant pas mis i en relation avec j.
Par convention, on ne compare jamais i avec lui-même.

Pour la comparaison (i,j) deux juges sont en accord si pour la case (i,j) des deux tableaux de comparaisons les valeurs sont les mêmes, sinon ils sont en désaccord.
Dans ces conditions le nombre de désaccords mutuels D_{ij} entre les m juges concernant la comparaison (i,j) est égal à $c_{ij}\,(m-c_{ij})$
Le nombre d'accords mutuels A_{ij} est égal à:

$$\frac{c_{ij}(c_{ij}-1)}{2} + \frac{(m-c_{ij})(m-c_{ij}-1)}{2}$$

et l'on a:

$$A_{ij} + D_{ij} = \frac{m(m-1)}{2}$$

Remarquons que $D_{ij}=0$ est possible. Cette valeur est obtenue lorsqu'il y a unanimité sur la comparaison (i,j); par contre, la valeur $A_{ij}=0$ est impossible.

Appel ons D le nombre total de désaccords mutuels et A le nombre total d'accords mutuels pour l'ensemble des comparaisons (i,j) au nombre de n(n-1).
On a:

$$D = \sum_{i \neq j} c_{ij}\,(m-c_{ij})$$

et:

$$A = \sum_{i \neq j} \left(\frac{c_{ij}(c_{ij}-1)}{2} + \frac{(m-c_{ij})(m-c_{ij}-1)}{2} \right)$$

Avec bien entendu:

$$A+D = \frac{n(n-1)\ m(m-1)}{2}$$

On aura donc:

$$A_{max}+D_{min} = A_{min}+D_{max} = \frac{n(n-1)\ m(m-1)}{2}$$

D'où: $A_{max}-A_{min} = D_{max}-D_{min}$

A_{max}, A_{min}, D_{max}, D_{min} représentant respectivement le nombre d'accords mutuels maximum et minimum et le nombre de désaccords mutuels maximum et minimum.

Coefficient d'accord mutuel "a"
Ce coefficient est défini par:

$$a = \frac{A-A_{min}}{A_{max}-A_{min}} = \frac{D_{max}-D}{D_{max}-D_{min}}$$

avec:
a=0 si $A=A_{min}$ ou $D=D_{max}$
a=1 si $A=A_{max}$ ou $D=D_{min}$

Coefficient de désaccord mutuel "d"
Ce coefficient est défini par:

$$d = \frac{D-D_{min}}{D_{max}-D_{min}} = \frac{A_{max}-A}{A_{max}-A_{min}}$$

Ces coefficients sont toujours définis et a+d=1 par construction.

Comme $D_{min}=0$ la valeur de d se réduit à:

$$d = \frac{D}{D_{max}} \qquad \text{et} \qquad a = 1 - \frac{D}{D_{max}}$$

Le Max de D est obtenu lorsque les valeurs c_{ij} et $(m - c_{ij})$ sont les plus voisines possible. Ce maximum est égal à:

$$\left[\frac{m}{2}\right]\left[\frac{m+1}{2}\right]$$

valeur valable quelle que soit la parité de m.

On a donc:

$$D_{max} = n(n-1) \left[\frac{m}{2}\right] \left[\frac{m+1}{2}\right]$$

D'où:

$$d = \frac{\sum_i \sum_{j \neq i} c_{ij} (m - c_{ij})}{n(n-1) \left[\frac{m}{2}\right] \left[\frac{m+1}{2}\right]}$$

On peut également définir ces coefficients a et d dans le cas de comparaisons manquantes.
Dans ce cas, on apppellera c_{ij} le nombre de juges ayant effectivement mis i en relation avec i et m_{ij} le nombre de juges ayant effectivement comparé i à j, $m_{ij} \leq m$. Le nombre de juges n'ayant pas mis i en relation avec j est donc $m_{ij} - c_{ij}$.
On aura alors:

$$d = \frac{\sum_i \sum_{j \neq i} c_{ij} (m_{ij} - c_{ij})}{\sum_i \sum_{j \neq i} \left[\frac{m_{ij}}{2}\right] \left[\frac{m_{ij}+1}{2}\right]}$$

On peut également donner une autre version de d en conservant la valeur du dénominateur initial, mais dans ce cas la signification de d est différente et le maximum de d peut être inférieur à 1, etc...

Mais cette amélioration, par rapport au coefficient U de Kendall, n'est pas suffisante. Même si un coefficient de "concordance" varie de 0 à 1, son intérêt dans le cadre d'un problème d'agrégation reste très limité.
Ceci pour au moins deux raisons:
1) le défaut essentiel d'un coefficient de concordance est que sa définition ne dépend que des tableaux de comparaisons individuels C^k et est indépendante de la valeur du tableau collectif Y. Un tel coefficient ne tient même pas compte de la nature du tableau collectif Y.
Nous considérons ici toujours le cas d'agrégation de m tableaux, de comparaisons par paires, par un tableau collectif particulier. Et nous avons vu que l'agrégation que nous considérons est une agrégation au sens large, ce que certains auteurs appelleraient selon le cas agrégation ou ajustement. Nous ne supposons pas que les tableaux individuels sont nécessairement de même nature que le tableau collectif. C'est une situation que l'on rencontre fréquemment dans la pratique, par exemple en sport ou en psychologie où des tableaux non transitifs sont à agréger

par un ordre total.
Si l'on prend comme exemple m tableaux de préférences portant sur trois objets A, B, C on peut très bien avoir A<B, B<C, et A<C à l'unanimité. Dans ce cas, le coefficient de concordance est égal à 1. Or ces tableaux n'étant pas transitifs, aucun ordre collectif Y ne pourra représenter parfaitement les tableaux individuels.
M.G. KENDALL est d'ailleurs conscient de cette lacune. Il a en effet indiqué dans [5] un troisième coefficient, dzeta, qu'il appelle "coefficient de consistance", mesurant les "inconsistances" (circuits de longueur 3) dans un tableau non transitif.
Mais ceci ne résoud pas le problème.
2) Un autre défaut de ces coefficients est d'être invariants à une "inversion des opinions" pour les comparaisons (i,j) et (j,i). C'est à dire une transformation du tableau C où c_{ij} devient c_{ji} et réciproquement laisse inchangée la valeur du coefficient de concordance, que ce soit le coefficient U ou a ou les variantes mentionnées précédemment.
Même si l'on se restreint au cas où les tableaux individuels et collectifs sont de même nature, des ordres totaux par exemple, cette invariance conduit à des résultats difficilement acceptables.
Prenons par exemple deux cas d'agrégation de classements, avec dans chaque cas 7 juges classant 3 objets A,B,C.

Exemple 1	Exemple 2
A A A B B C C	A A A B B C C
B B B A C A A	B B B C C A A
C C C C A B B	C C C A A B B

Les tableaux de comparaisons C correspondants ont pour valeur:

	A	B	C
A		5	4
B	2		5
C	3	3	

	A	B	C
A		5	3
B	2		5
C	4	2	

Ces deux tableaux sont identiques à l'exception des valeurs des cases AC et CA qui se trouvent inversées dans le deuxième tableau. Les deux tableaux ont donc le même coefficient de concordance.
Par exemple:

$$U = \frac{-1}{63} \quad \text{et} \quad a = \frac{1}{9}$$

bien qu'au niveau de l'agrégation il s'agisse de deux exemples très différents.

En effet, si nous prenons A B C comme classement collectif, le tableau Y le représentant vérifie la règle de la majorité pour chaque comparaison, c'est à dire on aura:

$$y_{ij} = 1 \quad \text{si } c_{ij} > \frac{m}{2}$$

$$y_{ij} = 0 \quad \text{si } c_{ij} < \frac{m}{2}$$

pour chaque comparaison (i,j); le tableau collectif Y est donc "très satisfaisant.
Par contre, le même tableau Y ne vérifie pas toujours la règle de la majorité pour les différentes comparaisons dans l'exemple 2.
En effet, dans cet exemple on a un effet Condorcet, c'est à dire A est préféré à B à la majorité de même que B à C et C à A; et Y ne vérifie pas la règle de la majorité pour la comparaison AC. Ce qui n'est "pas très satisfaisant".
De toute façon, aucun classement collectif ne peut être "très satisfaisant" puisque la présence d'un effet Condorcet interdit à tout classement de vérifier la règle de la majorité pour l'ensemble des comparaisons.
Il semble très souhaitable de disposer d'un coefficient pouvant indiquer cette différence de situation.
Les deux points précédents m'amènent à proposer un coefficient de nature différente du coefficient de concordance. Il s'agit d'un coefficient de "cohérence", indiquant la qualité d'une agrégation.

- Coefficient de cohérence

Les règles d'agrégation que nous avons considérées ont toujours la forme suivante:
l'opinion collective maximise un critère d'agrégation; et dans le cas où l'opinion collective représente un classement donné sous forme de rangs, ou un tableau de comparaisons par paires particulier, il s'agit d'une maximisation sous contraintes.
C'est cette dernière situation que nous allons considérer.
Comme précédemment, nous nous limiterons au cas d'agrégation de tableaux. Mais le principe reste valable dans le cas où les opinions sont des classements donnés sous forme de rangs; dans ce cas les résultats sont d'ailleurs beaucoup plus simples.

Appellons F(Y) la valeur du critère d'agrégation pour l'opinion collective Y. Par la suite, le critère

d'agrégation utilisé sera le "critère de la majorité", c'est à dire:

$$F(Y) = \sum_{i \neq j} c'_{ij} \; y_{ij}$$

où c'_{ij} = le nombre de juges ayant mis i en relation avec j moins le nombre de juges n'ayant pas mis i en relation avec j.

c'_{ij} est défini même dans le cas de comparaisons manquantes, mais ce principe est valable avec un critère quelconque.
Appellons F_{max} et F_{min} respectivement le maximun de F(Y) et le minimum de F(Y) pour Y sans contraintes, c'est à dire: y_{ij} =0 ou 1.
La cohérence c d'un tableau Y relativement au critère F sera défini par:

$$c = \frac{F(Y) - F_{min}}{F_{max} - F_{min}}$$

L'incohérence i sera définie par:

$$i = \frac{F_{max} - F(Y)}{F_{max} - F_{min}}$$

La structure de ces coefficients est analogue aux coefficients a et d. Les expressions A_{max}, A_{min} et A étant remplacées respectivement par F_{max}, F_{min} et F(Y).

On a: c+i=1

Ces coefficients sont toujours définis, sauf dans le cas où

$$F_{max} = F_{min}$$

Dans ce cas, F(Y) a une valeur constante indépendante de Y. Ce qui correspond au cas de non discrimination totale du critère F.

Lorsque F représente le critère de la majorité, les coefficients c et i sont donnés par:

$$c = \frac{\sum_{i \neq j} c'_{ij} \; y_{ij} - \sum_{i \neq j} \mathrm{Min}(0, c'_{ij})}{\sum_{i \neq j} |c'_{ij}|}$$

$$i = \frac{\sum_{i \neq j} \mathrm{Max}(0, c'_{ij}) - \sum_{i \neq j} c'_{ij} \; y_{ij}}{\sum_{i \neq j} |c'_{ij}|}$$

Puisque:

$$F_{max} - F_{min} = \sum_{i \neq j} \{Max(0, c'_{ij}) - Min(0, c'_{ij})\}$$
$$= \sum_{i \neq j} |c'_{ij}|$$

Voyons maintenant l'utilisation de ce coefficient dans le cadre d'une agrégation.
Dans un problème d'agrégation donné, un tableau collectif Y est d'un type particulier. Il doit vérifier un certain nombre de contraintes ou de relations que nous appellerons R. Par exemple, si Y représente un ordre total, Y sera total, antysimétrique et transitif. Et par définition, la règle d'agrégation donne un tableau collectif maximisant F(Y) pour Y vérifiaut R. Donc par construction la règle d'agrégation donne un tableau Y ayant le meilleur coefficient de cohérence c possible.
Mais bien entendu, Max F(Y) pour Y vérifiant R peut très bien être strictement inférieur à Max F(Y) pour Y sans contraintes (y_{ij}=0 ou 1), cette dernière valeur étant en quelque sorte "idéale", et dans ce cas on aura $c < 1$.

Nous considérons que pour le critère d'agrégation choisi, l'agrégation sera d'autant meilleure que la valeur c se rapproche de 1.
Pour le critère d'agrégation choisi nous considérons que l'agrégation sera d'autant meilleure que la valeur Max F(Y) pour Y vérifiant R se rapproche de la valeur dite "idéale" Max F(Y) pour Y sans contraintes (y_{ij}=o ou 1).
c =1 ou i = 0 lorsqu'il y a égalité et $c < 1$ ou i>0 sinon.
Par exemple, lorsque toutes les opinions représentent des ordres totaux, la valeur i = 0 correspond à l'absence d'effets Condorcet et i 0 à leur présence, la valeur de i indiquant en quelque sorte l'intensité de cet effet Condorcet.

Reprenons les deux exemples donnés précédemment et considérons dans chaque cas le tableau Y représentant le classement collectif ABC.

On vérifiera que pour ces 2 exemples le classement collectif ABC correspond à celui obtenu avec la règle de la majorité sous contraintes. On a donc avec ce classement collectif la meilleure valeur possible pour les coefficients i et c.

Dans le premier cas on a i= 0
Dans le 2ème cas on a i= 1/7

Ce qui n'est pas étonnant puisque l'exemple 1 ne contient pas d'Effet Condorcet tandis que l'exemple 2 en contient un.

On constate que l'agrégation de l'exemple 2 est moins bonne que celle de l'exemple 1, ce que ne peuvent indiquer les coefficients de concordance.

On peut généraliser cette approche de la façon suivante:

au lieu de considérer des tableaux sans contraintes, on s'interesse à des tableaux vérifiant des contraintes R_1 et l'on s'interesse à des tableaux collectifs vérifiant des contraintes R plus fortes que R_1.

La définition de i et c sera la même que précedemment, à la modification près suivante:

on remplacera Fmax =MaxF(Y) pour Y sans contraintes par:
F_{max} =MaxF(Y) pour Y vérifiant R_1

de même: Fmin =MinF(Y) pour Y sans contraintes
devient: F_{min} =MinF(Y) pour Y vérifiant R_1

Il s'agit en quelque sorte d'une cohérence ou incohérence "conditionnelle". Par exemple, si R_1 représente les conditions d'antisymétrie et de totalité, calculer la cohérence ou l'incohérence d'un classement Y relativement à R_1 revient à "mesurer l'incohérence" créee par la seule adjonction de la transitivité; alors que dans le cas initial, on mesurait l'incohérence créee à la fois par l'antysimétrie, la totalité et la transitivité.
Bien sûr R_1 peut être quelconque; et dans ce cas le calcul de F_{max} et F_{min} peut être relativement compliqué. Mais bien souvent dans la pratique R_1 a une forme simple et le calcul devient élémentaire.

Par exemple R_1 peut représenter les relations suivantes:

1) $y_{ij} + y_{ji} = 1$ antisymétrie et totalité
2) $y_{ij} + y_{ji} \geq 1$ totalité
3) $y_{ij} + y_{ji} \leq 1$ antisymétrie
4) $y_{ij} - y_{ji} = 0$ symétrie

Les valeurs correspondantes de F_{max} sont données par:

1) $\mathrm{Max}(c'_{ij}, c'_{ji})$
2) $\mathrm{Max}(c'_{ij}, c'_{ji}, c'_{ij} + c'_{ji})$
3) $\mathrm{Max}(c'_{ij}, c'_{ji}, 0)$
4) $\mathrm{Max}(c'_{ij} + c'_{ji}, 0)$

et l'on a des formules analogues pour F_{min} .

D)- REFERENCES

[1] K.ARROW,"Social Choice and Individual Value", 2ème édition, Wiley, New-York, (1963), 1ère édition (1951).

[2] J.C. BORDA, "Mémoire sur les Elections au Scrutin", Histoire de l'Académie Royale des Sciences 1781, Paris,(1784), Traduction Anglaise A. de Grazia, Isis, 44, (1953).

[3] A. de CONDORCET,"Essai de l'application de l'Analyse à la Probabilité des décisions rendues à la Pluralité des voix", Paris, (1785), réimpression Chelsea Publ. 6, New-York, (1974).

[5] M.G. KENDALL, "Rank Correlation Methods", 3ème edition, Hafner, New-York, (1962), 1ère edition (1948).

[6] F.MARCOTORCHINO, "Agrégation des Similarités en Classification Automatique", Thèse d'Etat présentée à l'Université Paris VI, (1981).

[7] B. MONJARDET, "Relations à Eloignement Minimum de Relations Binaires", Note Bibliographique, Math. et Sciences Humaines No 67, pp115-122, (1979).

P1

OPTIMIZATION IN ORDINAL DATA ANALYSIS

P. MICHAUD

F. MARCOTORCHINO

IBM FRANCE, PARIS SCIENTIFIC CENTER

This paper has been presented at the international symposium on Discrete Optimization August 1977, University of British Columbia, VANCOUVER ; with one exception, Chapters 5 and 6, which have been added in this second edition.

First edition January 1978

Second augmented edition October 1978

1. INTRODUCTION

Today's collection of data analysis solving methods is extremely important. Numerous methods of analysis, classification and structuration of data have been developed and implemented in many directions but not all. For instance the areas of ordinal data analysis and cluster aggregation seem to have been neglected in this respect.

However, ordinal data analysis is a subject which has been theoretically studied since a long time; of historical importance is a problem of voting studied by the Marquis Antoine de Condorcet in 1785.

Ordinal data analysis is also a subject of practical interest. Many real life problems encountered in economics, marketing, politics, sociology and in general in human sciences have to be solved by ordinal data analysis and especially by aggregation methods.

Finally the limitation of the use of ordinal data analysis is due particularly to the lack of practical solving methods. The same thing could be said about the cluster aggregation problem.

These aggregation problems can be stated as follows; individual preferences being given (most of the time contradictory) find a collective choice reflecting as much as possible all the individual preferences.

The first edition was limited to the study of aggregation problems encountered in ordinal data analysis; they were rating, ranking and opinion aggregation problems.

In this second edition, two new chapters have been added. They deal with other aggregation problems; the cluster aggregation problem and the general problem of binary relation aggregation.
We propose here various methods allowing to solve rapidly these different aggregation problems.
Non linear programming is used to solve the rating problem, network optimization to solve the ranking problem, while large scale linear programming is needed in the other cases.
Using the IBM MPSX/370 system and other improved optimization algorithms, we have built a conversational system in PL/1 under CMS.

2. THE RATING AGGREGATION PROBLEM

2.1. PRESENTATION OF THE PROBLEM

This first aggregation problem has the most quantitative and the simplest structure of all the types of aggregation problems presented in this paper.

In this problem several individual ratings of objects are given. The problem is to replace all the individual ratings by a single collective rating. This collective rating must be as close as possible of all the individual ratings. For instance ,the search of a collective rating of pupils accordings to the notes given by a certain number of teachers, finding a collective rating of firms according to their results on different financial ratios are problems of such a type.

The data are presented as follows:

The n "objects" and the m "ratings" are represented by a rating tableau T. In this tableau the kth column T_k represents the values of the kth rating for the n objects; t_{ik} is the value of the kth rating for the ith object.

In practice these ratings correspond to the notes or the values (those values must be positive number) obtained by the n objects on m different criteria.

These criteria can represent: either judges, examiners etc.., or economic variables, caracteristics etc.. .

The objects can represent: either candidates , students etc.., or elements of any statistical population etc.. .

In this kind of multidimensional configuration, the criteria

are often conflicting; and it is difficult to have an unidimensional point of view of the problem. The search of a collective rating is an answer to this question.

2.2. THE MODEL

We look for a collective rating vector $x = (x_1, \ldots\ldots, x_n)$ which is as close as possible of the other ratings T_k, $k=1,\ldots\ldots,m$. The proximity between the collective rating x and a rating T_k can be represented by a distance $d(x, T_k)$; for instance the Holder metric

$$d_\ell(x, T_k) = \sqrt[\ell]{\sum_{i=1}^{n} |x_i - t_{ik}|^\ell}$$

When the caracteristic ℓ tends to 0 or to infinity, the distance is still defined; we obtain

$$d_0(x, T_k) = \sum_{i=1}^{n} \mathrm{Sgn}|x_i - t_{ik}|$$

where Sgn is the sign function

$$d_\infty(x, T_k) = \max_i |x_i - t_{ik}|$$

The "distance" between a collective rating x and all the other ratings T_k, $k=1,\ldots\ldots,m$ can be defined by

$\sum_{k=1}^{m} d_\ell(x, T_k)$ (First model) or $\sqrt[\ell]{\sum_{k=1}^{m} (d_\ell(x, T_k))^\ell}$ (second model)

the unknown collective rating is then solution of

$$(1) \qquad \min \sum_{k=1}^{m} \sqrt[\ell]{\sum_{i=1}^{n} |x_i - t_{ik}|^\ell}$$

in the first model and

$$(2) \qquad \text{Min} \sqrt[\ell]{\sum_{k=1}^{m} \sum_{i=1}^{n} |x_i - t_{ik}|^{\ell}}$$

in the second model.

For the particular values 0 and infinity of the caracteristic ℓ, only one representation of the objective function is used. It is given by:

$$\text{Min} \sum_{k=1}^{m} \sum_{i=1}^{n} \text{Sgn}|x_i - t_{ik}| \quad \text{for } \ell = 0$$

$$\text{Min} \max_{k} \max_{i} |x_i - t_{ik}| \quad \text{for } \ell = \infty$$

For the general values of the caracteristic ℓ, these two non linear (convex) minimization problems, (1) and (2), are not equivalent at all; the structure of the objective function and of the optimal solution are very different:

1) The objective function

In the first model the objective function is not separable (a function is separable if it is the sum of functions of one variable). the partial derivatives are not defined for $x = T_k$. The resolution of this problem is nevertheless

possible but requires the use of a general algorithm for non linear unconstrained programs; for instance the conjugate gradients method of Fletcher and Reeves or the variable metric method of Davidon Fletcher and Powell,see[8] and [7]. The second model is much easier; the objective function is separable. In general the iterative Newton method with one variable can be used. Even more, in some cases, the value of the optimal solution can be given explicitely.

2) The optimal solutions

In the first model (1) , related to the Fermat-Weber's problem, the optimal solution has no simple geometric or statistical interpretation, even in the simplest classical case of the Fermat's problem, which corresponds to the use of the euclidean metric.

In the second model (2), the solutions have a very interesting interpretation in terms of classical statistical quantities. For these reasons we shall concentrate upon the second model, although the resolution of the first model is possible by using either the conjugate gradients method or the variable metric method.

3) Resolution (of the second model)

Since the objective function is convex and separable for $1 < \ell < \infty$, the obtainment of the optimal solution consists in solving the following equation:

$$f_i'(x_i) = 0 \quad \text{with} \quad f_i(x_i) = \sum_{k=1}^{m} | x_i - t_{ik} |^{\ell}$$

In the general case $1< \ell <\infty$ the (prime) derivative is given by :

$$f_i'(x_i) = \sum_{k=1}^{m} \ell\,(x_i - t_{ik})\,|x_i - t_{ik}|^{\ell-2}$$

and the second derivative is given by:

$$f_i''(x_i) = \sum_{k=1}^{m} \ell(\ell-1)\,|x_i - t_{ik}|^{\ell-2}$$

Then for the resolution of the previous equation the Newton's method can be used. The optimal solution is obtained in that case by sequential iterations with a quadratic convergence. The iterative formula is given by:

$$x_i^{t+1}(\ell) = x_i^t(\ell) - \frac{f_i'(x_i^t(\ell))}{f_i''(x_i^t(\ell))}$$

In our case we obtain :

$$x_i^{t+1}(\ell) = \frac{\ell - 2}{\ell - 1}\, x_i^t(\ell) + \frac{\sum_{k=1}^{m} t_{ik}\,|x_i^t(\ell) - t_{ik}|^{\ell-2}}{(\ell-1)\sum_{k=1}^{m} |x_i^t(\ell) - t_{ik}|^{\ell-2}}$$

In the case where the caracteristic ℓ of the Holder metric is equal to 1 and 2 , the optimal solution can be given explicitely,for instance :

If $\ell=1$, the ith component x_i of the optimal rating vector x is equal to the Median of the distribution t_{ik} for k=1,2,..m

If $\ell=2$, the ith component x_i of the associated optimal solution is equal to the Mean of the t_{ik}'s for k=1,2..m

Moreover the two particular limit cases with $\ell=0$ and $\ell = \infty$ have also explicit solutions:

If $\ell=0$, the ith component of the associated optimal rating is

the Mode of the distribution t_{ik} for k=1,2..m

If $\ell = \infty$, the ith component of the associated optimal rating is the Center (of the Range) of the distribution t_{ik} for k=1,2,..m.

We can see here that the optimal solutions have an interesting interpretation in terms of statistical quantities. In general, the optimal solution is a function of the classical statistical quantities both of central tendancy and of deviation more and more complex with the increase of ℓ . For instance for the case $\ell=4$, the ith component of the optimal solution is a function of the Mean of the Standard Deviation and of the coefficient of skweness of the distribution t_{ik} for k=1,2...m:

$$x_i^*(4) = \left(\sum_{k=1}^{m} t_{ik}\right)/m + \sigma_i F(\alpha_i)$$

Where σ_i is the standard deviation of the distribution and $F(\alpha_i)$ is a function of the coefficient of skewness

4) Improvement of the modelization

In practice some transformations must be done , in some cases, on the data before initiating the optimization phase; for instance:

If some t_{ik} are < 0 , we must make them positive by translations.

If the units are too different we must harmonize the data by linear transformations.

In the case where the original ratings after transformations are still ill-conditionned, we have the possibility to use

other approaches:

a) Use of weights

Whithout changing the properties of the previous objective function, we can introduce the concept of weight to the model. In other words it is possible to associate with each rating T_k a weight w_k corresponding to the importance given to the rating. The objective functions become :

$$\sum_{k=1}^{m} w_k \left(d_\ell (x , T_k) \right)^\ell$$

b) Use of the Khi-square metric d_{χ^2}

In that case we want to harmonize the profiles of each rating in order to make the data homogeneous either by weighting the rows or by weighting the columns. It is possible in that case to use a particular metric which is not really a distance but which gives in practice very good results, the khi-square metric d_{χ^2} . With this metric the proximity between two ratings is given by:

$$d_{\chi^2}(T_k , T_\ell) = t_{..} \sum_{i=1}^{n} \frac{1}{t_{i.}} \left[\frac{t_{ik}}{t_{.k}} - \frac{t_{i\ell}}{t_{.\ell}} \right]^2$$

$$t_{..} = \sum_{i,k} t_{ik} \quad \text{(sum over all the values)}$$

$$t_{i.} = \sum_{k} t_{ik} \quad \text{(sum over the row } i \text{)}$$

$$t_{.k} = \sum_{i} t_{ik} \quad \text{(sum over the column k)}$$

The associated objective function,computed from the augmented tableau $T \cup x$, is then given by:

$$\text{Min} \sum_{k=1}^{m} \sum_{i=1}^{m} \frac{t_{..} + \sum_j x_j}{t_{i.} + x_i} \left[\frac{x_i}{\sum_j x_j} - \frac{t_{ik}}{t_{.k}} \right]^2$$

As we can see this objective function is strongly non linear, and for obtaining the optimal solution , we can use one of the general optimization methods defined previously.

3. THE RANKING AGGREGATION PROBLEM

3.1. PRESENTATION OF THE PROBLEM

We have seen in the previous chapter, how to transform raw data into ratings or scales.

But this transformation is not always meaningful, even if we use the Khi-square metric, particularly if the criteria are too heterogeneous.

To avoid this restriction we can replace each individual rating or scale by its associated induced ranking ; in other words each rating represented by the column T_k must be replaced by a ranking represented by the column R_k , (as in the aggregation problem, ties are possible). This is the ranking aggregation problem.

But in that case the expected collective result must be a complete ranking $x = (x_1 , x_2 , \ldots x_n)$ summarizing the m individual rankings R_k. This new constraint : the fact that x must be a complete ranking (i.e a permutation) has to be added in order to keep homogeneity as far as rankings are concerned.

Thus the expression of the associated objective function in the general case is still given by:

$$\sum_{k=1}^{m} w_k \, (d_\ell \, (x , R_k))^\ell$$

for the Holder d_ℓ metrics

with the additional constraint:

$$x \in S_n$$

where S_n is the symmetric group of the n permutations, and where w_k represent the weights, as in the previous Chapter associated to the m criteria.

Remark:

For the particular Khi-square metric , the expression of the objective function is

$$\sum_{k=1}^{m} w_k \, d_{\chi^2}(x, R_k)$$

As we have already seen, the general associated optimization model for the Holder metric is

$$\text{Min} \sum_{k=1}^{m} w_k \, (d_\ell(x, R_k))^\ell$$

$$x \in S_n$$

This new aggregation problem is no longer an unconstrained problem because of the permutation constraint and it is more complex a priori than the rating aggregation problem.

3.2. LINEARIZATION OF THE PROBLEM

If we keep this formulation of the problem, (non linear objective function, non linear constraint and integer variables), the problem is very difficult even impossible to solve.

To avoid this difficulty we have proposed transformations of

the formulation of this problem; the result of these transformations is the replacement of the original problem by a linear one.

We shall see that this hard problem can be transformed in a linear one.

3.2.1.Linearization of the permutation constraint

We can replace the unknown ranking vector $x=(x_1, x_2, \ldots x_n)$, by a binary permutation matrix X, representing the complete ranking x.

The general term x_{ij} of this matrix has the following signification:

$x_{ij} = 1$ if $x_i = j$, i.e. the rank of the object i is j in x

$x_{ij} = 0$ otherwise

Since the constraint $x \in S_n$ is expressed in terms of a permutation matrix, we have the following relations equivalent to the previous constraint.

$$\sum_{i=1}^{n} x_{ij} = 1 \qquad j=1,2,\ldots,n$$

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i=1,2,\ldots,n$$

3.2.2.Linearization of the objective function

The unknown ranking x_i of the object i in the vector x , can be expressed in terms of x_{ij}'s in the following way:

$$x_i = \sum_{j=1}^{n} j\, x_{ij}$$

We have shown, see [13], that the replacement of x by its new

formulation in the previous optimization model gives the unexpected following result: the objective function becomes linear in x_{ij}, and the global problem becomes a network problem for all the metrics used in the first chapter, (in general a linear assignment problem except for the d_∞ metric for which the problem is then a bottleneck assignment problem).

This formulation is now a classical one, allowing to solve quickly problems of very large size.

After all these transformations the rankings aggregation problems can be written as follows:

$$\text{Min} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}(d,w)\, x_{ij}$$

$$\sum_{i=1}^{n} x_{ij} = 1 \qquad j=1,2....n$$

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i=1,2....n$$

$$x_{ij} \geq 0$$

Where the quantity $c_{ij}(d\ ,\ w\ )$ represents the cost corresponding to the assignment of the object i to the rank j.

3.2.3.The assignment costs

This cost c_{ij} depends simultaneously on the m rankings given by the judges, on the chosen metric and if they are used on the weights associated to the judges.

The computation of the c_{ij}'s for all the metrics is detailed in [13].

In order to illustrate how this result can be obtained, we will consider the case of the khi-square metric, a similar

approach will do for the other cases.

In that case we have:

$$d_{\chi^2}(x, R_k) = \frac{2(m+1)}{n(n+1)} \sum_{i=1}^{n} \frac{(x_i - r_{ik})^2}{x_i + r_{i.}}$$

where

$$r_{i.} = \sum_{k=1}^{m} r_{ik}$$

The expression of d_{χ^2} is simpler than in the case of the first Chapter, since the sum of the ranks given by a judge k must be a constant equal to $\frac{n(n+1)}{2}$.

The objective function of the associated optimization problem is given by:

$$\sum_{k=1}^{m} w_k \; d_{\chi^2}(x, R_k)$$

If we assume that $w_k = 1 \;\; \forall$ k, for more simplification, and if we replace $d_{\chi^2}(x, R_k)$ by its value in terms of the r_{ik}'s, we obtain:

$$\frac{2(m+1)}{n(n+1)} \sum_{i=1}^{n} \sum_{k=1}^{m} \frac{(x_i - r_{ik})^2}{x_i + r_{i.}}$$

Since we have seen that:

$$x_{ij} = 1 \quad \text{if } x_i = j$$
$$x_{ij} = 0 \quad \text{otherwise}$$

we can express the objective function under the following linear form in terms of x_{ij}'s:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left[\sum_{k=1}^{m} \frac{(j - r_{ik})^2}{j + r_{i.}} \right] x_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}(d_{\chi^2}) \; x_{ij}$$

If $x_{ij} = 1$, i.e $x_i = j$, the term between brackets is taken into account , its value corresponds to its contribution in the objective function.

If $x_{ij} = 0$, this term is not taken into account, that completes the proof of the linearity of the objective function.

But we can see here, that the cost $c_{ij}(d_{\chi^2}, w)$ is not a linear function of the w_k's, because these weights appear in the denominator in the quantity $r_{i.} = \sum_k w_k r_{ik}$.

However for the general case of an Holder metric, we have a stronger result. Indeed the expression of the cost c is given by:

$$c_{ij}(d_\ell , w) = \sum_{k=1}^{m} w_k \mid j - r_{ik} \mid^\ell$$

w_k represents the weight associated to the judge k

r_{ik} represents the rank of the object i in the ranking of the judge k

ℓ represents the caracteristic of the chosen Holder metric.

We can see here that c_{ij} is a linear function of the weights w, and that c_{ij} does depend on the m individual rankings and on the metric.

As previously the optimal solution x, depends on the choice

of the metric. More similar the optimal rankings are, for the different metrics, more significant the results are.

3.3. VALIDATION OF THE OPTIMAL SOLUTION

For providing us against discrepancies in the results , we have two checking facilities.

The first one , applicable a priori, concerning the data, is the use of concordance indexes measuring rankings skewness.
The second , which is an a posteriori approach, consists in a test of the sensibility and of the validity of the solutions to the weights variation.
These two facilities are complementary to each other, and can be used independantly.

3.3.1. Concordance indexes

In order to measure if the search of the expected result is meaningful according to the data structure, we have defined a concordance index, both for the Holder metric and for the Khi-square metric. This index is a generalization of the index given by M.G.Kendall in the case of the Euclidean metric [11]. In the case of the Holder metric, this index is a direct extension of the Kendall's index,while for the Khi-square metric this index is based upon the Khi-square test and upon the entropy (according to the Shannon's definition),(see [13]).
This index is computed a priori from the previous ranking

tableau R, and indicates if the given data are coherent enough for an eventual research of a solution.

In the general case of an Holder d_ℓ metric, this concordance index is a function of the given m rankings and also a function of the number m of judges; we obtain (see [13]):

$$[\, W_\ell(m)\,]^\ell = \frac{\sum_{i=1}^{n} \left| r_{i.} - \frac{m(n+1)}{2} \right|^\ell}{(m^\ell/2^{\ell-1}) \sum_{i=1}^{n} |n+1 - 2i|^\ell}$$

This index $W_\ell(m)$ can vary from 0 to 1:

If the rankings given by the m judges are not very different one from the other, the value of $W_\ell(m)$ is close to 1.

If ,on the opposite , the judges give opinions without concordances the value of $W_\ell(m)$ is close to 0. This index is then a good measure of the rankings skewness.

If we consider a collective ranking x, one can measure its coherence with the other given rankings by using the concordance index W; if now this collective ranking is unknown, we have the opportunity to find directly the one that maximizes the concordance index for the augmented tableau $R \cup x$, (i.e maximizing the collective agreement).

In that case, we have shown , (see[13]), that we obtain also a linear assignment problem; and in general for a given metric, the optimal solutions derived from the minimization of the previous objective function and from the maximization of the concordance indexes are not equivalent.

From W_ℓ (m), we can derive the value $W_\ell(m+1, x)$ for the previous R tableau plus the unknown added ranking x .
We obtain:

$$[\, W_\ell(m+1,x)\,]^\ell = \frac{\sum_{i=1}^{n} \left| r_{i.} + x_i - \frac{(m+1)(n+1)}{2} \right|^\ell}{\frac{(m+1)^\ell}{2^{\ell-1}} \sum_{i=1}^{n} |n+1 - 2i|^\ell}$$

Since the denominator is a constant, the problem:

$$\max_{x} \; [\, W_\ell(m+1,\, x\,)\,]$$

$$x \in S_n$$

can be written as follows:

$$\text{Max} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} \left| r_{i.} + j - (m+1)\frac{n+1}{2} \right|^\ell x_{ij}$$

$$\sum_{i=1}^{n} x_{ij} = 1 \qquad j = 1,2....n$$

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i = 1,2....n$$

$$x_{ij} \geqslant 0$$

3.3.2.Sensibility to the weights

We have seen previously that the main goal in rankings aggregation is to find a collective ranking, but this problem is not the only one. In particular two different problems can be added to this.

a) The search of the stability of an optimal solution

In other words, we can check whether an optimal solution remains unchanged, when we modify the value of one of the

weights.

b) The validation of an a priori collective ranking

In other words, we can verify whether it is possible to find a distribution of weights , allowing to obtain as optimal solution of the assignment problem, the ranking given a priori.

These tests are especially easy to work out in the case where we are dealing with costs $c_{ij}(d, w)$ linear in terms of w_k 's;(i.e for the Holder metric).
In that case these two problems admit a resolution by linear programming; we will describe them briefly below.

All the optimal solutions of the following assignment problem, where the weights w_k are given:

$$\text{Min} \sum_i \sum_j c_{ij}(d_\ell, w)\, x_{ij}$$

$$\sum_i x_{ij} = 1 \quad j=1,2....n$$

$$\sum_j x_{ij} = 1 \quad i=1,2....n$$

$$x_{ij} \geq 0$$

satisfy optimality conditions. In order to state these optimality conditions it is necessary to introduce the dual variables (u_i and v_j), associated to the constraints of the previous problem, we obtain:

$$c_{ij}(d_\ell, w) - (v_j - u_i) = 0 \quad \text{for} \quad x_{ij} = 1$$

$$c_{ij}(d_\ell, w) - (v_j - u_i) \geq 0 \quad \text{for} \quad x_{ij} = 0$$

However, since we are in the case of an Holder metric the costs c_{ij} (d_ℓ, w) are linear in w_k . If we replace the costs by their linear expression in terms of w_k 's, we obtain a new formulation of the optimality conditions:

$$\sum_{k=1}^{m} w_k \mid j - r_{ik} \mid^{\ell} - (v_j - u_i) = 0 \text{ if } x_{ij} = 1$$

$$\sum_{k=1}^{m} w_k \mid j - r_{ik} \mid^{\ell} - (v_j - u_i) \geqslant 0 \text{ if } x_{ij} = 0$$

From these optimality conditions, we can derive the statement of the two previous problems.

a) Sensibility of an optimal solution to the variation of a weight

In that case, we use the optimality conditions in the following way:

We can isolate the term corresponding to the weight we want to modify and put in the right-hand side the constant terms associated to all the other unchanged weights. We can eliminate also the v_j 's (or the u_i's), and the n equality constraints. If we assume, moreover, that the variable weight must be non negative, we obtain now a system of $n^2 - n + 1$ inequalities with n+1 variables.

In other words, if after modification the weight still verifies the previous system, the optimal ranking will remain unchanged. We will look for the maximum then the minimum

values of this weight, verifying these constraints. This will give the bounds.

These bounds are obtained by solving two linear programming problems with the following objective functions:

$$\text{Min (or Max)} \quad w_k$$

under the previous inequalities considered as constraints for the L.P. problem.

The difference between the maximum and the minimum values $w_{max} - w_{min}$ represents the range. If $w_{max} - w_{min} = 0$ the solution is not stable, because every variation of this chosen weight will modify the optimal solution.

This approach is applicable for all the weights. A judge plays a prominent part in the optimal collective ranking if the variation of its associated weight is very small.

b) Validation of a given solution

Let us assume we are given an a priori collective ranking, we wonder if it is possible to find a distribution of weights, allowing to obtain as optimal solution of the previous assignment problem, the a priori ranking .

To do so, we will use the previous optimality conditions, but in that case all the weights can vary. This leads to another system of $n^2-n+m+1$ inequalities with $n+m$ variables; the n

equations and the v_{m_j} variables are eliminated and $w_k \geq 0$ $k=1,\ldots,m$, and $\sum_{k=1} w_k \geq 1$ must be added. If there exist a combination of the m weights verifying this system, the ranking given a priori is compatible with the preferences of the judges. The resolution of this system of inequalities will be done by linear programming.

4. THE OPINION AGGREGATION PROBLEM

4.1. PRESENTATION OF THE PROBLEM

This third class of aggregation problem has a more qualitative, and also a more complex structure than the two previous classes.

The theoretical study of the subject has been taken up many times, in numerous articles, but the associated practical solving methods have been missing until now.

This opinion aggregation model is a generalisation of the previous problems and at the same time a different model since it allows to take into account data which are neither quantitative nor ordinal but qualitative.

The presentation of the individual data is also different of the ones given previously. In this problem the point of view of a judge k is represented by a pairwise comparison tableau C^k , also called opinion tableau. This is a(0,1) tableau where (the case i=j is never considered) :

$$\begin{cases} c^k_{ij} = 1 & \text{if the object i is preferred to the object j for the judge k.} \\ c^k_{ij} = 0 & \text{otherwise.} \end{cases}$$

The opinion of a judge k can also be represented by an alternative comparison tableau C'^k . This is a(-1,+1)tableau where:

$$\begin{cases} c'^k_{ij} = 1 & \text{if the object i is preferred to the object j for the judge k.} \\ c'^k_{ij} = -1 & \text{otherwise.} \end{cases}$$

The expression of c'^k_{ij} in term of c^k_{ij} is $c'^k_{ij} = 2\, c^k_{ij} - 1$.

Any rating or ranking, for a judge k, can be represented by a comparison tableau, $c^k_{ij} = 1$ if $t_{ik} > t_{jk}$ for a rating or if $r_{ik} < r_{jk}$ for a ranking, but the converse is not necessarily true; the comparison tableau can represents opinions which do not correspond to an order.

The important concept of opinion in term of pairwise comparison is due to the Marquis Antoine de Condorcet who introduced this term in 1785 for a voting problem.

Summing up all the individual tableaux C^k we obtain a collective tableau C which represents the opinions of all the group of judges;

$c_{ij} = \sum_k c^k_{ij}$. Each c_{ij} represents the number of judges preferring i to j

The alternative collective tableau C' is the sum of all the individual tableaux C'^k , $c'_{ij} = \sum_k c'^k_{ij}$.

Each c'_{ij} represents the number of judges preferring i to j minus the number of judges not preferring i to j.

When every judge gives a complete ranking $c_{ij} + c_{ji} = m$, the number of judges and $c'_{ij} + c'_{ji} = 0$; in that case i not preferred to j is equivalent to j preferred to i. But this result is not true in general.

The opinion of the group of judges is given by the C tableau. The next step is to propose a collective ranking reflecting as much as possible all the opinions given.

This collective ranking is represented, like the C^k tableaux, by a binary tableau Y.

$y_{ij} = 1$ if i is ranked higher than j in the collective ranking,

$y_{ij} = 0$ otherwise.

As for the rating or the ranking problems we have to define:

a) A distance $d(C^k,Y)$ between Y and C^k.

b) A "distance" between Y and all the C^k.

a) Distance between Y and C^k

The Holder distance between the collective ranking Y and a C^k tableau is given by:

$$d_\ell^\ell (Y , C^k) = \sum_i \sum_j | y_{ij} - c_{ij}^k |^\ell$$

This distance has the following properties:

1) This distance is linear in term of Y

The linearity is due to the separability of the expression in term of Y. Let us consider an arbitrary separable function of Y. It can be written $f(Y) = \sum_i \sum_j f_{ij}(y_{ij})$.
But since y_{ij} is equal to 0 or 1 this former expression is equal to:

$$f(Y) = \sum_i \sum_j f_{ij}(0) + \sum_i \sum_j (f_{ij}(1) - f_{ij}(0))y_{ij}$$

and this is a linear expression in term of Y. This result applies to $d_\ell^\ell(\ C^k, Y\ )$ which is a separable function of Y; we obtain:

$$d_\ell^\ell(\ C^k,Y) = \sum_i \sum_j |c_{ij}^k|^\ell + \sum_i \sum_j [|1- c_{ij}^k|^\ell - |c_{ij}^k|^\ell]y_{ij}$$

2) All the Holder distances are independant of the caracteristic ℓ of the metric

Since c_{ij}^k is equal to 0 or 1, $|c_{ij}^k|^\ell$ and $|c_{ij}^k - 1|^\ell$ are equal to c_{ij}^k and $1 - c_{ij}^k$ respectively; and the expression of $d_\ell^\ell(C^k,Y)$ is equal to:

$$\sum_i \sum_j c_{ij}^k + \sum_i \sum_j (1 - 2\, c_{ij}^k)\, y_{ij}$$

Finally the expression of $d_\ell^\ell(C ,Y)$ is independant of ℓ and is given by

$$\sum_i \sum_j c_{ij}^k - \sum_i \sum_j c'^k_{ij}\, y_{ij}$$

Remark:

In the language of set theory $d_1(C^k, Y\ )$ represents the symmetric difference distance between two sets. The two sets are C^k and Y; their elements are represented by 1 in each tableau. The symmetric difference between two sets is equal to the number of elements in one set but not in the other; this is precisely the value given by

$$\sum_i \sum_j c_{ij}^k - \sum \sum c'^k_{ij}\, y_{ij}$$

b) Distance between Y and all the C^k

As in the case of the rating or the ranking problem, the distance between Y and all the C^k is represented by

$$\sqrt[\ell]{\sum_{k=1}^{m} d_\ell^\ell(Y, C^k)}$$

This is equivalent (same solution) to minimize the objective function $\sum_{k=1}^{m} d_\ell^\ell(Y, C^k)$

This objective function is independant of the caracteristic ℓ of the metric; it is equal to $\sum_k d_1(C^k, Y)$, for $\ell=1,2.....$, which was not the case for the rating or the ranking problem. The expression of this unique objective function is given by

$$\sum_i \sum_j \left(\sum_k c_{ij}^k\right) - \sum_i \sum_j \left(\sum_k c_{ij}^{!k}\right) y_{ij}$$

which is equal to:

$$\sum_i \sum_j c_{ij} - \sum_i \sum_j c_{ij}^{!} y_{ij}$$

Suppose that i is ranked higher than j, in the collective ranking Y, that is y_{ij} is set to one.

In that case, from the definition of C, we have c_{ij} agreements and c_{ji} disagreements; that is to say c_{ij} judges have also ranked i higher than j while c_{ji} judges have done the opposite choice.

Similarly, $c_{ij}^{!}$ represents the number of agreements minus the number of nonagreements; that is to say the number of judges having ranked i higher than j minus the number of judges not

having ranked i higher than j

If we sum up all the $c'_{ij} y_{ij}$, for all the couples (i,j), we obtain the total number of agreements minus the total number of nonagreements with the choice Y for the total ranking.

Finally the problem is to find the collective ranking Y which minimizes the distance between Y and all the C^k. This is equivalent to find the collective ranking Y maximizing the total number of agreements minus the total number of nonagreements. From the Y tableau, we deduce the optimal ranking; the rank r_i of the object i is given by:

$$r_i = \sum_{j \neq i} y_{ji} + 1 = n - \sum_{j \neq i} y_{ij}$$

But since Y represents a complete ranking the y_{ij} 's are not arbitrary; they must satisfy the following properties caracterising a complete ranking:

1) Completness (and anti-symmetry)

If we consider any two objects i and j we have only two exclusive possibilities; whether i is ranked higher than j or j is ranked higher than i.

The y_{ij}'s must satisfy

$$y_{ij} + y_{ji} = 1 \qquad \text{for } i = 1,\ldots\ldots n \text{ , } j \neq i.$$

2) Transitivity

Let us consider any three objects i, j and k. If we have i ranked higher than j and also j ranked higher than k in the final ranking then i must be ranked higher than k.

In other words $y_{ij} = 1$ and $y_{jk} = 1$ imply $y_{ik} = 1$. This is equivalent to the following relation

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1.$$

4.2. THE ASSOCIATED MODEL

Our problem, to find a collective ranking maximizing the total number of agreements minus the total number of nonagreements, can be represented by a linear programming model with zero-one variables.

This programm is defined as follows:
The zero-one variables are the y_{ij} values of the ranking Y.
The objective function, to maximize, represents the total number of agreements minus the total number of nonagreements.
The constraints represent the transitivity and the completness conditions defined previously. These constraints insure that the Y tableau is a complete ranking.
Finally, we obtain the following model

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij}\, y_{ij}$$

$$y_{ij} + y_{ji} = 1 \qquad i \neq j$$

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1 \qquad i \neq j,\ j \neq k,\ i \neq k$$

$$y_{ij} = 0 \text{ or } 1$$

This model has n(n-1) relations of completness (and antisymmetry) and n(n-1)(n-2) relations of transitivity.
This model can be reduced to (see [13] p. 139):

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij}\, y_{ij}$$

$$y_{ij} + y_{ji} = 1 \qquad i < j$$

$$0 \leqslant y_{ij} + y_{jk} - y_{ik} \leqslant 1 \qquad i < j < k$$

$$y_{ij} = 0 \text{ or } 1$$

With this time $n(n-1)/2$ relations of completness and $n(n-1)(n-2)/6$ double relations of transitivity.

Remarks:

1) This problem is equivalent to the maximization of the total number of agreements, $\sum_{ij\neq i} \sum c_{ij} y_{ij}$. Since Y is a total order $\sum_{i} \sum_{j\neq i} y_{ij} = \frac{n(n-1)}{2}$ and $\sum_{ij\neq i} \sum c'_{ij} y_{ij} = 2\sum_{ij\neq i} \sum c_{ij} y_{ij} - m\frac{n(n-1)}{2}$

2) This problem is also equivalent to the minimization of the total number of disagreements $\sum_{i} \sum_{j\neq i} c_{ji} y_{ij}$ and of course to the maximization of the total number of agreements minus the total number of disagreements $\sum_{i} \sum_{j\neq i} (c_{ij} - c_{ji}) y_{ij}$. (See [13], p.137)

3) Any of these objective functions can be divided by $\sum_{i} \sum_{j} c_{ij}$; this gives normalized objective functions equivalent to the original ones.

4) When the opinion tableaux do not represent rankings even the case of a single opinion tableau is of interest. This is the classical problem of graph theory where a binary relation, the comparaison tableau C , must be approximated by an order relation, the Y tableau.
A well known exemple of this is the tournament problem where, in the comparison tableau C , $c_{ij} = 1$ if i dominates j and $c_{ij} = 0$ otherwise .

4.2.1.The associated linear programming problem.

If we relax the binary condition on the variables, we obtain a linear programming problem given by:

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij}\, y_{ij}$$

$$(1) \qquad y_{ij} + y_{ji} = 1 \qquad i<j$$

$$(2) \qquad 0 \leqslant y_{ij} + y_{jk} - y_{ik} \leqslant 1 \qquad i<j<k$$

$$(3) \qquad 0 \leqslant y_{ij} \leqslant 1$$

Since (1) and $y_{ij} \geqslant 0$ imply $y_{ij} \leqslant 1$, (3) can be replaced by

(3') $0 \leqslant y_{ij}$. This problem can be rewritten in standard or inequality forms.

a)- The standard form

We have to introduce here new variables y_{ijk} and y_{kji} ; y_{ijk} is defined by $y_{ijk} = y_{ij} + y_{jk} - y_{ik}$,while y_{kji} is a slack variable.

If we introduce these new variables in the previous program we obtain:

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij}\, y_{ij}$$

$$(i) \qquad y_{ij} + y_{ji} = 1 \qquad i<j$$

$$(ii) \qquad y_{ijk} + y_{kji} = 1 \qquad i<j<k$$

$$(iii) \qquad y_{ij} + y_{jk} - y_{ik} - y_{ijk} = 0$$

$$0 \leqslant y_{ij} \qquad i<j$$

$$0 \leqslant y_{ji}$$

$$0 \leqslant y_{ijk} \qquad i<j<k$$

$$0 \leqslant y_{kji}$$

This program has $n' = \frac{(n+1)n(n-1)}{3}$ variables and $m' = \frac{(2n-1)n(n-1)}{6}$ relations; the size of the bases is m'.

b)- The inequality form

In the previous model, we eliminate the variables y_{ji} for i<j. Each double relation (2) and (3) is replaced by two single relations.

We obtain

$$\text{Max} \sum_i \sum_{i<j} (c'_{ij} - c'_{ji}) y_{ij} + K$$

$$\begin{aligned} - y_{ij} - y_{jk} + y_{ik} &\leqslant 0 \quad \text{with dual variable } u'_{ijk} \\ y_{ij} + y_{jk} - y_{ik} &\leqslant 1 \quad " \quad " \quad " \quad u''_{ijk} \end{aligned} \qquad i<j<k$$

$$\begin{aligned} - y_{ij} &\leqslant 0 \quad \text{with dual variable } u'_{ij} \\ y_{ij} &\leqslant 1 \quad " \quad " \quad " \quad u''_{ij} \end{aligned}$$

With: $K = \sum_{i>j} c'_{ij}$

This primal program is in inequality form, its dual is in standard form and is given by:

$$\text{Min} \left[\sum_{i<j<k} u''_{ijk} + \sum_{i<j} u''_{ij} \right] + (K)$$

$$\sum_{j<\ell} (u''_{ij\ell} - u'_{ij\ell}) + \sum_{\ell<i} (u''_{\ell ij} - u'_{\ell ij}) - \sum_{i<\ell<j} (u''_{i\ell j} - u'_{i\ell j}) + u''_{ij} - u'_{ij} = c'_{ij} - c'_{ji} \qquad i<j$$

$$\begin{aligned} 0 &\leqslant u'_{ijk} \\ 0 &\leqslant u''_{ijk} \end{aligned} \qquad i<j<k$$

$$\begin{aligned} 0 &\leqslant u'_{ij} \\ 0 &\leqslant u''_{ij} \end{aligned} \qquad i<j$$

The primal has $n' = \frac{n(n-1)}{2}$ variables and $m' = \frac{(n+1)n(n-1)}{3}$ relations. These two quantities are reversed for the dual problem. The bases of the primal and of the dual are identical; their size is equal to the number n' of variables of the primal problem.

Therefore it is advisable to use the inequality form rather than the standard form to solve the linear program.

4.2.2. Relation between the bivalent model and the linear program

In general the bivalent model and the associated linear model are not equivalent. In particular:

a) The matrix of the constraints defining these two problems is not unimodular; this fact contradicts a result given by V. J. Bowman in [4] pp. 586 and 587.

b) Each basic feasible solution has not necessarily zero-one coordinates, though that was implicitely assumed by de Cani in [5].

Since all the vertices of the polyhedron, of the linear program, are not in zero-one, it is possible to built an economic function, for which the optimal solution is located on a non integer vertex of the polyhedron; and the optimal solutions of the two problems do not coincide. (At this occasion, we would like to thank E.Johnson for providing us with an example with n=6). We have also shown that for any

integer tableau C there exists always another C tableau different of the original but equivalent(same optimal solution) that is the sum of opinions tableaux C^k. This means that from a theoretical point of view, the possibility of fractional values for an opinion problem exists.

However for a practical problem the c_{ij}'s are not arbitrary. In particular the integers c_{ij} must satisfy

$$0 \leqslant c_{ij} + c_{jk} - c_{ik} \leqslant m \qquad \text{and}$$

$$c_{ij} + c_{ji} = m.$$

In the case where the individual data represent total rankings.

Thus in the case of total rankings, it is possible to wonder if the possibility of obtaining fractional values really exists; but this is just a conjecture.

4.3. THE PROPOSED METHOD

The proposed method is a linear programming method caracterised by the two following points:

a) Resolution of the model with the MPSX/370 system.

b) Use of the linear program under inequality form.

These two caracteristics are extremly important since they will allow us to solve real life problems of large size ; this is what we are precising now:

a) Use of the MPSX/370 system (IBM mathematical programming system)

It is necessary to use such a system for the following reasons:

1) The MPSX/370 system possesses features such as LU decomposition of the basic matrices, Devex method to select the variables entering and leaving the basis, handling of degeneracy cases etc... see[2] and [15].

These techniques, either original or recent, allow us to solve large problems in a particularly efficient way.

2) Moreover, this system possesses an extended control language (ECL) in PL/1 language [15]. It is then possible to use this system in a program written in PL/1. In such a program this system seems to be a PL/1 subroutine. With such an approach the generation of the linear program and the interpretation of the results is especially easy; which is not always the case with other control languages.

b) Use of the linear program under inequality form

We have seen that this inequality form leaded to bases of the order of n squared instead of n cubed. This allows the resolution of much bigger problems.

As far as the resolution is concerned, the existing codes consider always the reference problem under standard form; a problem not initially under standard form is made standard by automatic adjunction of slack variables. Two options are possible for the resolution: resolution of the reference

problem or resolution of the dual of the reference problem. These two options are called respectivelly "primal procedure" and "dual procedure". Since the reference problem is turned in standard form its dual appears under inequality form. In both cases the size of the bases is the same; it is the number of constraints of the reference problem.

For this reason, one must not work with the original problem under inequality form . The automatic adjunction of slack variables to make it standard would lead to bases of the order of n cubed; which will remove all interest to the approach. It is the dual of this problem which must be taken as reference problem: it is originally under standard form with bases of size $\frac{n(n-1)}{2}$.

Thus the "dual procedure" of the MPSX/370 system will allow the resolution of the original problem under inequality form; the "primal procedure" the resolution of the dual of the original problem. This does not represent any difficulties; the generator used being able to generate indifferently the original problem or its dual.

We have used the "primal procedure" (resolution of our dual problem) to solve numerous real life problems, about a hundred. In all the cases, we have obtained an optimal solution with zero-one coordinates.

We give here some results concerning the computing time obtained with an IBM 370/168 computer using the MPSX/370 system.

This computing time includes the reading of the data, the

generation of the linear program, the resolution of the problem and the editing of the results in term of rankings.

- For most of the problems, those with $10 < n < 30$ the computing time was from 3 to 20 seconds.
- For a bigger problem, with $n = 36$, the computing time was 30 seconds.
- For a much bigger problem, with $n = 72$, an optimal solution was obtained in less than 15 minutes; which is relatively a short time, considering the size of the linear programming problem(2556 constraints, 124392 variables and 427705 non zero elements in the constraint matrix). It is the first time, at least to our knowledge, that an opinion problem of such a size has been solved by a method giving the optimal solution.

In principle the "dual procedure" is less powerful; the time of a "dual" iteration is in general longer than the one of a "primal" iteration.

However this "dual procedure" is to be recommended in our case. Since we solve directly the initial problem, under inequality form, it becomes possible to initiate the procedure with a starting solution defined beforehand; which is not the case of the "primal procedure" working on the dual problem.

We use as starting solution, the solution obtained with a very efficient heuristic we have proposed in [13]. This starting solution is in general very close to the optimal solution (and often equal, but the heuristic is unable to

prove the optimality).

We have to generate a basis associated with this solution as a starting basis.

If we proceed in this way, we obtain a considerable reduction of the number of iterations; this is expressed by a net gain in the computing time, all the more important as the time of the iterations is preponderant compaired to the total computing time.

5. THE SIMILARITY AGGREGATION PROBLEM

5.1. PRESENTATION OF THE PROBLEM

This fourth type of aggregation problem is no longer an ordinal data analysis problem; the collective tableau and the individual data are not related to rankings but to partitions (a partition is a collection of disjoint clusters). It is not an ordinal data analysis problem but simply a data analysis problem. This does not mean that this problem is simpler; on the contrary this (combinatorial) problem is even more complex than the opinion problem as far as its structure is concerned.

The presentation of the data is the same as the one for the opinion problem, but the meaning is different.

In this problem, the point of view of a judge k is represented by a comparison tableau S^k, called similarity tableau. This is a (0,1) tableau where:

$s^k_{ij} = 1$ if the object i is similar to the object j for the judge k

$s^k_{ij} = 0$ otherwise.

The opinion of a judge k can also be represented by an alternative comparison tableau S'^k.

This is a (-1,+1) tableau where:

$s'^k_{ij} = 1$ if the object i is similar to the object j for the judge k

$s'^k_{ij} = -1$ otherwise.

The expression of s'^k_{ij} in term of s^k_{ij} is $s'^k_{ij} = 2\, s^k_{ij} - 1$

Any partition, for the judge k, can be represented by a

comparison tableau, $s^k_{ij} = 1$ if i and j are in the same class (cluster), but the converse is not necessarily true.

The comparison tableau can represent similarities which do not correspond to partitions. This arises in various applications when the S^k tableau is built from a proximity tableau P^k; the proximity tableau is not necessarily symmetric and $s^k_{ij} = 1$ if p^k_{ij} is less than a given threshold, $s^k_{ij} = 0$ otherwise.

Summing up all the individual tableaux S^k we obtain a collective tableau S which represents the similarities expressed by all the group of judges; $s_{ij} = \sum_k s^k_{ij}$. Each s_{ij} represent the number of judges considering i similar to j.

The alternative collective tableau S' is the sum of all the individual tableaux S'^k ; $s'_{ij} = \sum_k s'^k_{ij}$. Each s'_{ij} represents the number of judges considering i similar to j minus the number of judges not considering i similar to j.

When every judge gives a partition $s_{ij} = s_{ji}$ and $s'_{ij} = s'_{ji}$; in that case, i similar to j is equivalent to j similar to i. But this result is not true in general.

The next step is to propose a collective partition reflecting as much as possible all the similarities given.

This collective partition is represented, like the S tableau, by a binary tableau Y.

$y_{ij} = 1$ if i is similar to j in the collective partition

$y_{ij} = 0$ otherwise .

But in this case i similar to j means i is in the same class as j.

The distance between two tableaux is the one used for the opinion problem.

The distance between S^k and Y is

$$d_1(Y, S^k) = \sum_i \sum_{j \neq i} s^k_{ij} - \sum_i \sum_{j \neq i} s'^k_{ij} y_{ij}$$

The "distance" between Y and all the S^k is

$$\sum_i \sum_{j \neq i} s_{ij} - \sum_i \sum_{j \neq i} s'_{ij} y_{ij}$$

Like in the opinion problem, minimizing this distance is equivalent to maximize :

$$\sum_i \sum_{j \neq i} s'_{ij} y_{ij}$$

Let us suppose that i is in the same class as j, in the collective partition, that is to say y_{ij} is set to one.
In that case, from the definition of S, we have s_{ij} agreements; that is to say s_{ij} judges have considered i similar to j (the notion of disagreement is not defined for this problem).
Similarly, s'_{ij} represents the number of agreements minus the number of "nonagreements"; that is to say the number of judges having considered i similar to j minus the number of

judges not having considered i similar to j.

If we sum up all the $s'_{ij}\, y_{ij}$, for all the couples (i,j), we obtain the total number of agreements minus the total number of nonagreements with the choice Y for the partition.

Finally the problem is to find a collective partition Y which minimizes the distance between Y and the S^k . This is equivalent to find a collective partition Y maximizing the total number of agreements minus the total number of nonagreements.

From the Y tableau, we deduce the optimal partition; i and j are in the same class if $y_{ij} = 1$.

Since Y represents a partition, the y_{ij}'s are not arbitrary; they must satisfy the following properties caracterising a partition:

1) Symmetry

Let us consider two objects i and j; if i is in the same class as j then j is in the same class as i.

The y_{ij}'s must satisfy

$$y_{ij} = y_{ji} \quad \text{i.e.} \quad y_{ij} - y_{ji} = 0 \qquad \text{for } i = 1,\ldots,n \quad j \neq i$$

2) Transitivity

This property has been detailed in the opinion problem. For this problem, it is expressed as follows:

If i is in the same class as j and j is in the same class as k, then i is in the same class as k.

The y_{ij}'s must satisfy

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1.$$

5.2. THE ASSOCIATED MODEL

As in the case of the opinion problem, the similarity problem can be represented by a linear program with zero-one variables.

The zero-one variables are the y_{ij} values of the collective partition Y.

The objective function, to maximize ,is $\sum_i \sum_{j \neq i} s'_{ij} y_{ij}$

The constraints represent the transitivity and the symmetry conditions defined previously. They insure that the Y tableau represents a partition.

We obtain the following model:

$$\text{Max} \sum_i \sum_{j \neq i} s'_{ij} y_{ij}$$

$$y_{ij} - y_{ji} = 0 \qquad i \neq j$$

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1 \qquad i \neq j \; j \neq k \; i \neq k$$

$$y_{ij} = 0 \text{ or } 1$$

This model has $n(n-1)$ relations of symmetry and $n(n-1)(n-2)$ relations of transitivity. As in the opinion

problem the relations can be simplified, and this model can be reduced to :

$$\text{Max} \sum_i \sum_{j \neq i} s'_{ij} y_{ij}$$

$$y_{ij} - y_{ji} = 0 \quad i<j$$

$$-y_{ij} + y_{jk} + y_{ik} \leqslant 1$$

$$y_{ij} - y_{jk} + y_{ik} \leqslant 1 \quad i<j<k$$

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1$$

$$y_{ij} = 0 \text{ or } 1$$

With this time $\frac{n(n-1)}{2}$ relations of symmetry and $\frac{n(n-1)(n-2)}{6}$ triple relations of transitivity.

Remarks:

1) Unlike in the opinion problem, Max $\sum_i \sum_{j \neq i} s'_{ij} y_{ij}$ is not equivalent to Max $\sum_{ij \neq i} s_{ij} y_{ij}$. This later problem has an obvious optimal solution; all the objects are in the same class.

2) The similarity tableaux can represent partitions. In this particular case the problem is known under the name of "Central Partition Problem" see [12] and [16].

3) When the similarity tableaux do not represent partitions, even the case of a single similarity tableau is of interest. This is the problem of graph theory where a binary relation,

the similarity tableau S, must be approximated by an equivalence relation, the Y tableau.

A particular example of this is the Zahn's problem where a symmetric relation is approximated by an equivalence relation, see [17].

5.2.1.The associated linear programming problem

If we relax the binary condition on the variables, we obtain a linear programming problem given by:

$$\text{Max} \quad \sum_i \sum_{j \neq i} s'_{ij} \, y_{ij}$$

$$(1) \qquad y_{ij} - y_{ji} = 0 \qquad i<j$$

$$(2) \qquad \begin{aligned} -y_{ij} + y_{jk} + y_{ik} &\leqslant 1 \\ y_{ij} - y_{jk} + y_{ik} &\leqslant 1 \\ y_{ij} + y_{jk} - y_{ik} &\leqslant 1 \end{aligned} \qquad i<j<k$$

$$(3) \qquad 0 \leqslant y_{ij} \leqslant 1$$

a)-The standard form

We can eliminate simultaneously the symmetry relations and the y_{ji} variables for $i<j$ from the previous program. We introduce new slack variables $z_{ijk}, z'_{ijk}, z''_{ijk}$ and z_{ij} for the relations (2) and (3) in the previous program. We obtain:

$$\text{Max} \sum_i \sum_{i<j} \left(s'_{ij} + s'_{ji} \right) y_{ij}$$

$$\begin{aligned}
-y_{ij} + y_{jk} + y_{ik} + z_{ijk} &= 1 \\
y_{ij} - y_{jk} + y_{ik} + z'_{ijk} &= 1 \qquad i<j<k \\
y_{ij} + y_{jk} - y_{ik} + z''_{ijk} &= 1
\end{aligned}$$

$$\begin{aligned}
y_{ij} + z_{ij} &= 1 \\
y_{ij} &\geqslant 0 \qquad i<j \\
z_{ij} &\geqslant 0 \\
z_{ijk} &\geqslant 0 \\
z'_{ijk} &\geqslant 0 \qquad i<j<k \\
z''_{ijk} &\geqslant 0
\end{aligned}$$

This program has $n'=n^2(n-1)/2$ and $m'=n(n-1)^2/2$; the size of the bases is m'.

b)-<u>The inequality form</u>

In the original linear program, we just eliminate the y_{ji} for $i<j$.

We obtain:

$$\text{Max} \sum_i \sum_{i<j} (s'_{ij} + s'_{ji})\, y_{ij}$$

$$\begin{aligned}
-y_{ij} + y_{jk} + y_{ik} &\leqslant 1 \text{ with dual variable } u_{ijk} \\
y_{ij} - y_{jk} + y_{ik} &\leqslant 1 \text{ with dual variable } u'_{ijk} \qquad i<j<k \\
y_{ij} + y_{jk} - y_{ik} &\leqslant 1 \text{ with dual variable } u''_{ijk}
\end{aligned}$$

$$\begin{aligned}
y_{ij} &\leqslant 1 \quad " \quad " \quad " \quad u_{ij} \qquad i<j \\
-y_{ij} &\leqslant 0 \quad " \quad " \quad " \quad u'_{ij}
\end{aligned}$$

This primal program is in inequality form, its dual is in standard form and is given by

$$\text{Min} \sum_{i<j<k} (u_{ijk} + u'_{ijk} + u''_{ijk}) + \sum_{i<j} u_{ij}$$

$$\sum_{\ell<i} (u_{\ell ij} - u'_{\ell ij} + u''_{\ell ij}) + \sum_{i<\ell<j} (u_{i\ell j} + u'_{i\ell j} - u''_{i\ell j}) + \sum_{j<\ell} (-u_{ij\ell} + u'_{ij\ell} + u''_{ij\ell}) + \cdots$$

$$\cdots + (u_{ij} - u'_{ij}) = s'_{ij} + s'_{ji} \qquad i<j$$

$$\begin{aligned} u_{ijk} &\geq 0 \\ u'_{ijk} &\geq 0 \quad i<j<k \\ u''_{ijk} &\geq 0 \end{aligned}$$

$$\begin{aligned} u_{ij} &\geq 0 \\ u'_{ij} &\geq 0 \end{aligned} \quad i<j$$

The primal has $n'=n(n-1)/2$ variables and $m'=n^2(n-1)/2$ relations. These two quantities are reversed for the dual problem. The bases of the primal and of the dual are identical; their size is equal to the number n' of variables of the primal problem.

5.3. THE PROPOSED METHOD

The proposed method is identical to the one used for the opinion problem; we solve the associated (continuous) linear

program under inequality form, using the MPSX/370 system.
The only modification concerns the generator; the constraints are slightly different of those of the opinion problem. The number of the constraints is $n^2 (n-1)/2$ for this problem compared to $n(n+1)(n-1)/3$ for the opinion problem.
The number of variables and the size of the bases is unchanged.

The results are similar to those obtained with the opinion problem:

- The polyhedron of the zero-one model and of the associated linear program are not in principle identical.
- However we have always reached an optimal solution with zero-one coordinates for all the practical problems we have solved (about 20 at the time of this writing).
- The computing time (with the same computer as before) was from 5 to 20 seconds for problems with $10<n<30$. A central partition problem with $n=35$ has been solved in less than 90 seconds.

6. THE RELATION AGGREGATION PROBLEM

6.1. PRESENTATION OF THE PROBLEM

The opinion and similarity aggregation problems are just particular cases of relation aggregation problems.
In the first case, binary relations are approximated by a complete order relation. In the second case, binary relations are approximated by an equivalence relation.
Complete orders and partitions have a common property; they correspond to particular transitive relations (quasi-orders). But there exists other kinds of transitive relations.

The general problem of relation aggregation consists precisely of approximating binary relations by these other transitive relations.

The transitive relations can be divided into six groups:

1) A transitive relation can be just a transitive relation (quasi-order).
2) When a transitive relation (noted $\leqslant$) is at the same time symmetric, that is to say $i \leqslant j$ implies $j \leqslant i$, it is an equivalence relation.
3) When a transitive relation is at the same time anti-symmetric, that is to say $i \leqslant j$ implies $j \leqslant i$ $(i \neq j)$, it is an order relation.

For each of these three cases we have two possibilities; the relation can be partial or complete.
A partial relation corresponds to the general case.
A relation is complete when at least one of the two expressions $i \leqslant j$ or $j \leqslant i$ holds for every i and j ($i \neq j$).

One of these groups, the group of complete equivalence relations, is of no interest and will not be considered here; it corresponds to the unique case where all the elements are equivalent and are in the same cluster.

Of the five remaining cases, two have already been studied (the complete order and the equivalence relations), the other cases are:

- -Partial transitive relation
- -Complete transitive relation
- -Partial order relation

Approximating binary relations by one of those relations can be done exactly as in the case of complete order or equivalence relations.

The notations used are those of the approximation by a complete order relation (opinion problem), see p. 24 .
The C^k and C'^k tableaux represent each of the k binary relations.
$C = \sum_k C^k$ and $C' = \sum_k C'^k$ are the "collective" tableaux.

The Y tableau represents the particular transitive relation to find.

The "distance" between Y and all the other C^k tableaux is

$$\sum_{ij\neq i} \sum c_{ij} - \sum_{i} \sum_{j\neq i} c'_{ij}\, y_{ij}$$

The Y tableau satisfies the transitivity conditions.

If Y represents a symmetric relation, the y variables must satisfy $y_{ij} = y_{ji}$ for $i<j$.

If Y represents a antisymmetric relation, the y_{ij} variables must satisfy $y_{ij}+y_{ji} \leqslant 1$ for $i<j$.

If Y represents a complete relation, the y_{ij} variables must satisfy $y_{ij}+y_{ji} \geqslant 1$ for $i<j$.

In each case, the problem is to find a particular transitive relation Y which minimizes the distance between Y and the C^k. Again, this problem is equivalent to maximize $\sum_{i} \sum_{j\neq i} c'_{ij} y_{ij}$ the total number of agreements minus the total number of nonagreements.

6.2. THE ASSOCIATED MODEL

As in the case of the opinion and similarity problems, the relation agregation problem can be represented by a linear program with zero-one variables.

The variables are the y_{ij} values of Y.

The objective function, to maximize, is $\sum_{i} \sum_{j\neq i} c'_{ij}\, y_{ij}$

The constraints represent the transitivity conditions plus the conditions defined previously.

We obtain the following models:

1) Model for a partial transitive relation

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij} y_{ij}$$

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1 \quad i \neq j,\ j \neq k,\ i \neq k$$

$$y_{ij} = 0 \text{ or } 1 \quad i \neq j$$

2) Model for a complete transitive relation

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij} y_{ij}$$

$$y_{ij} + y_{ji} \geqslant 1 \quad i<j$$

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1 \quad i \neq j, j \neq k, i \neq k$$

$$y_{ij} = 0 \text{ or } 1 \quad i \neq j$$

3) Model for a partial order relation

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij} y_{ij}$$

$$y_{ij} + y_{ji} \leqslant 1 \quad i<j$$

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1 \quad i \neq j, j \neq k, i \neq k$$

$$y_{ij} = 0 \text{ or } 1 \quad i \neq j$$

Unlike in the opinion and similarity problems the number of transitivity relations, $n(n-1)(n-2)$, cannot be reduced and the $n(n-1)/2$ variables y_{ij} with $i>j$ cannot be eliminated.
The three models have $n(n-1)$ variables.
The partial transitive relation model has $n(n-1)(n-2)$ constraints.
The complete transitive relation and the partial order relation have each $n^2(n-1)/2$ constraints.

6.2.1.The associated linear programming problem

If in the previous models, we replace the binary condition on the variables by $0 \leqslant y_{ij} \leqslant 1$, for $i \neq j$, we obtain linear programs.
If we add a slack variable to each inequality constraint, we obtain a linear program under standard form. But the order of the size is n cubed.
As in the opinion and similarity problems we have to consider the linear programming models under inequality form.
In that case the order of the size of the bases is n squared.
To obtain the linear programming models in inequality form, no extra constraints or variables is needed since each linear programming model is written only in terms of inequalities; $0 \leqslant y_{ij} \leqslant 1$ must be expressed as two inequalities $y_{ij} \leqslant 1$ and $-y_{ij} \leqslant 0$ for $i \neq j$.
As before the dual program is necessary for the computation.
The primal and the dual programs are given by:

1) Linear programs for the partial transitive relation problem.

- Primal:

$$\text{Max} \sum_i \sum_{j\neq i} c'_{ij}\, y_{ij}$$

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1 \qquad i\neq j, j\neq k, i\neq k$$

$$y_{ij} \leqslant 1 \qquad i\neq j$$

$$-y_{ij} \leqslant 0$$

- Dual:

$$\text{Min} \sum_i \sum_j \sum_k u_{ijk} + \sum_i \sum_{j\neq i} u_{ij} \quad \text{(with } i\neq j, j\neq k, i\neq k \text{ in the triple summation)}$$

$$\sum_{\substack{\ell\neq i \\ \ell\neq j}} u_{ij\ell} - \sum_{\substack{\ell\neq i \\ \ell\neq j}} u_{i\ell j} + \sum_{\substack{\ell\neq i \\ \ell\neq j}} u_{\ell ij} + u_{ij} - v_{ij} = c'_{ij} \qquad i\neq j$$

$$u_{ijk} \geqslant 0 \qquad i\neq j\neq k$$

$$u_{ij} \geqslant 0 \qquad i\neq j$$

$$v_{ij} \geqslant 0 \qquad i\neq j$$

u_{ijk} is associated to the transitivity constraint ijk

u_{ij} is associated to the constraint $y_{ij} \leqslant 1$

v_{ij} is associated to the constraint $-y_{ij} \leqslant 0$

2) Linear programs for the complete transitive relation problem.

- Primal:

$$\text{Max} \sum_{ij \neq i} \sum c'_{ij} \, y_{ij}$$

$$y_{ij} + y_{ji} \geq 1 \qquad i<j$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \qquad i \neq j, j \neq k, i \neq k$$

$$y_{ij} \leq 1 \qquad i \neq j$$

$$-y_{ij} \leq 0$$

- Dual:

$$\text{Min} \sum_i \sum_j \sum_k u_{ijk} + \sum_i \sum_{j \neq i} u_{ij} - \sum_i \sum_{i<j} w_{ij}$$

(with $i \neq j, j \neq k, i \neq k$ in the triple summation)

$$\sum_\ell u_{ij\ell} - \sum_\ell u_{i\ell j} + \sum_\ell u_{\ell ij} + u_{ij} - v_{ij} - w_{ij} = c'_{ij} \qquad i<j$$

$$\sum_\ell u_{ij\ell} - \sum_\ell u_{i\ell j} + \sum_\ell u_{\ell ij} + u_{ij} - v_{ij} - w_{ji} = c'_{ij} \qquad i>j$$

(with $\ell \neq i$, $\ell \neq j$ for these constraints)

$$u_{ijk} \geq 0 \qquad i \neq j \neq k$$

$$u_{ij} \geq 0 \qquad i \neq j$$

$$v_{ij} \geq 0 \qquad i \neq j$$

$$w_{ij} \geq 0 \qquad i<j$$

w_{ij} is associated to the constraint $y_{ij} + y_{ji} \geqslant 1$.

There are only $n(n-1)/2$ variables w_{ij}, but $n(n-1)$ variables u_{ij} and $n(n-1)$ variables v_{ij}.

3) Linear programs for the partial order relation problem.

- Primal:

$$\text{Max} \sum_{ij \neq i} c'_{ij}\, y_{ij}$$

$$y_{ij} + y_{ji} \leqslant 1 \qquad i<j$$

$$y_{ij} + y_{jk} - y_{ik} \leqslant 1 \qquad i \neq j, j \neq k, i \neq k$$

$$-y_{ij} \leqslant 0 \qquad i \neq j$$

- Dual:

$$\text{Min} \sum_i \sum_j \sum_k u_{ijk} + \sum_i \sum_{i<j} w_{ij}$$

(with $i \neq j, j \neq k, i \neq k$ in the triple summation)

$$\sum_\ell u_{ij\ell} - \sum_\ell u_{i\ell j} + \sum_\ell u_{\ell ij} - v_{ij} + w_{ij} = c'_{ij} \quad \text{for } i<j$$

$$\sum_\ell u_{ij\ell} - \sum_\ell u_{i\ell j} + \sum_\ell u_{\ell ij} - v_{ij} + w_{ji} = c'_{ij} \quad \text{for } i>j$$

(In the summations ℓ is different from i and j)

$$u_{ijk} \geqslant 0 \quad i \neq j \neq k$$

$$v_{ij} \geqslant 0 \quad i \neq j$$

$$w_{ij} \geqslant 0 \quad i<j$$

The u_{ij} variables do not appear in the dual problem. This is because the constraints $y_{ij} \leqslant 1$, $i \neq j$, have been eliminated in the primal problem; these constraints are implied by the constraints $y_{ij} + y_{ji} \leqslant 1$, $i<j$.

In the three cases the solving method will be the one defined for the opinion and the similarity problems.

7.REFERENCES

[1] ASTIE A., " Comparaisons par Paires et Problèmes de Classement" in Mathématiques et Sciences Humaines ,N 32,(1970).

[2] BENICHOU M., GAUTHIER J.M., HENTGES G. and RIBIERE G., " The Efficient Solution of Large Scale Linear Programming Problems. Some Algorithmic and Computational Results", in Mathematical Programming , Vol.13,N 3,280,(1977).

[3] BLIN J.M., " A Linear Formulation of the Multiattribute Decision Problem", in R.A.I.R.O. N6,21,(1976).

[4] BOWMAN V.J., " Permutation Polyhedra",in SIAM J.Appl.Math.,Vol.22, N4,580,(1972).

[5] de CANI J.S.," Maximum Likelihood Paired Comparison Ranking by Linear Programming",in Biometrica, Vol.56,N3,537,(1969).

[6] DAVIDON W.C.," Variable Metric Method for Minimization",in Argonne National Laboratory, A.N.L. 5990, (1959).

[7] FLETCHER R. and POWELL M.J.D.,"A Rapidly Convergent Descent Method for Minimization",in Computer Journal,Vol.6,163,(1963).

[8] FLETCHER R. and REEVES C.M.,"Function Minimization by Conjugate Gradients",in Computer Journal,Vol.7,149,(1964).

[9] FORD L.R. and FULKERSON D.R.,"Flows in Networks".Princeton,Princeton University Press, (1962).

[10] GROSS O., " The Bottleneck Assignment Problem",Santa Monica ,The Rand Corporation,P.1630,(1959).

[11] KENDALL M.G."Rank Correlation Methods".London,Griffin,(1962).

[12] LERMAN I.C.," Les Bases de la Classification Automatique".Paris, Gauthier-Villars,(1970).

[13] MARCOTORCHINO F.and MICHAUD P.,"Optimisation en Analyse Ordinale des Données".Paris,Masson (to appear December 1978).

[14] MONJARDET B.,"Tournois et Ordres Médians pour une Opinion",in Mathématiques et Sciences Humaines,N43,55,(1973).

[15]MPSX/370,"IBM Mathematical Programming System Extended/370",General Information Manual,GH19-1090.

[16]REGNIER S.,"Stabilité d'un opérateur de Classification",in Mathématiques et Sciences Humaines,N60,21,(1977).

[17]ZAHN C.T.,"Approximating Symmetric Relations by Equivalence Relations" SIAM J.Appl.Math.N12,840,(1964).

Math. Sci. hum. (17e année, n°67, *1979*, p.7-38)

P2

MODELES D'OPTIMISATION EN ANALYSE DES DONNEES RELATIONNELLES

Pierre MICHAUD François MARCOTORCHINO

1. INTRODUCTION

Le problème général que nous considérons dans cet article est le suivant: il s'agit de "l'agrégation" ou de "l'approximation" de plusieurs relations binaires quelconques par une relation binaire particulière.
Les problèmes tels que les problèmes d'agrégation de classements ou de préférences ainsi que les problèmes d'agrégation de partitions ou de similarités en sont des cas particuliers.

Dans le premier de ces deux cas, un certain nombre de relations binaires sont à approximer par une relation d'ordre total (classement). Dans le second cas elles sont à approximer par une relation d'équivalence (classification).

Dans le cas de l'agrégation de préférences (classement automatique), il n'existe à l'heure actuelle que des méthodes d'énumération implicite pour résoudre le problème de façon exacte; or ces méthodes sont très onéreuses en temps de calcul et pratiquement ne permettent pas de classer plus de 20 objets.

Dans le cas de l'agrégation de similarités (classification automatique) , il existe des algorithmes performants, comme la méthode des "nuées dynamiques" de E.Diday; mais cette méthode , comme les autres (en particulier l'algorithme des transferts de S.Régnier),n'est qu'une méthode heuristique ne garantissant qu'un optimum local du problème considéré.

Dans le présent article nous montrerons que pour ces problèmes, et d'autres problèmes voisins, il existe une modélisation linéaire.

Par cette approche linéaire, nous traiterons ces deux problèmes d'agrégation de préférences et d'agrégation de similarités qui sont les plus importants pour les applications pratiques. Nous généraliserons cette approche au cas de l'agrégation de relations binaires par d'autres types de relations, telles que les relations de préordre ou de quasi-ordre etc...

Cette modélisation repose sur les deux principes suivants:
1) La prise en compte des données sous forme relationnelle.
2) L'utilisation d'une règle de majorité.

En plus de la résolution des problèmes d'agrégation de préférences et de similarités, la linéarisation des modèles permet en outre d'obtenir la réponse à diverses questions ((3), (6), (11)) restées ouvertes.

2. PRESENTATION DU PROBLEME

Le problème général que nous traiterons dans le présent article est le suivant: il s'agit de "l'agrégation" ou de "l'approximation" de plusieurs relations binaires par une relation binaire particulière.
les problèmes évoqués dans l'introduction, tels que le problème de l'agrégation de préférences ou le problème de l'agrégation de similarités en sont des cas particuliers.

Dans le premier de ces deux cas, un certain nombre de relations binaires sont à approximer par une relation d'ordre total. Dans le second cas elles sont à approximer par une relation d'équivalence.
Les relations d'ordre total et d'équivalence ont en commun la propriété suivante: ce sont des relations transitives particulières.
Mais il existe d'autres types de relations transitives utilisables.

Le problème général de l'agrégation de relations binaires, que nous considérons, consiste précisemment à approximer des relations binaires par l'une ou l'autre de ces relations transitives particulières.
L'approximation de m relations binaires données par l'une de ces relations transitives sera réalisée par une approche générale valable, à quelques modifications près, pour l'ensemble des cas considérés.

3. PRESENTATION DES DONNEES

3.1. PRESENTATION RELATIONNELLE DES DONNEES

Chacune des m relations binaires de départ, définissant les données individuelles d'un "juge" ou "critère" numéro k, sera représentée par un tableau C^k. Ce tableau, que nous appellerons tableau d'opinion ou de comparaisons, est un tableau en (0,1) de terme général:

$c_{ij}^k = 1$ Si i "est en relation" avec j pour le juge k.
$c_{ij}^k = 0$ Sinon

Par convention la valeur c_{ij} avec i=j ne sera jamais considérée.

Suivant le type de problème à traiter et la nature des données à considérer, la signification des valeurs cij sera différente.

Par exemple:
-Si la relation binaire C^k représente les préférences individuelles d'un juge k, la signification du tableau C^k sera la suivante:

$c_{ij}^k = 1$ si l'objet i est préféré à l'objet j pour le juge k
$c_{ij}^k = 0$ sinon.

-Dans le cas où les tableaux C^k représentent des similarités entre différents objets, la signification du tableau C^k est la suivante:

$c_{ij}^k = 1$ si l'objet i est semblable ou similaire à l'objet j, pour le juge k.
$c_{ij}^k = 0$ sinon.

A chaque relation binaire représentée par un tableau C^k, on associe un autre tableau, noté C'^k ; ce nouveau tableau, à valeurs (-1,+1), est défini de la façon suivante:

$c'^k_{ij} = 1$ si l'objet i "est en relation " avec l'objet j, pour le juge k.
$c'^k_{ij} = -1$ sinon

l'expression de c'^k_{ij} en termes de c^k_{ij} est donnée par

$$c'^k_{ij} = 2\, c^k_{ij} - 1$$

L'expression de c^k_{ij} en termes de c'^k_{ij} est donnée par:

$$c_{ij}^{k} = (c_{ij}^{'k} + 1)/ 2$$

L'obtention des tableaux C^k ou C'^k peut se faire directement ou par l'intermédiaire de tableaux de données annexes.

En sommant tous les tableaux individuels C^k, représentant les relations binaires de départ, on obtient un tableau collectif C, qui représente l'opinion de l'ensemble des m juges.

On a donc : $C = \sum_k C^k$ et $c_{ij} = \sum_k c_{ij}^k$

Le terme général c_{ij} du tableau collectif C représente d'une façon générale le nombre de relations binaires où i est en relation avec j; c'est le nombre de juges ayant mis i en relation avec j.

De la même façon le tableau collectif C' est défini par:

$$C' = \sum_k C'^k \quad \text{et} \quad c'_{ij} = \sum_k c_{ij}^{'k} \ .$$

Le terme général c'_{ij} du tableau collectif C' est égal au nombre de juges ayant mis i en relation avec j moins le nombre de juges n'ayant pas mis i en relation avec j.

3.2. GENERALISATIONS POSSIBLES

- On peut combiner les tableaux collectifs entre eux. L'addition d'un tableau collectif C1, représentant l'opinion de m1 juges, et d'un tableau collectif C2, représentant l'opinion de m2 juges, donne un tableau collectif C , représentant l'opinion de m juges; avec C=C1+C2 et m=m1+m2.

- On peut diviser les tableaux C ou C' par m ,le nombre de juges. On obtient alors des tableaux de "fréquences".

Remarque: Ces tableaux de fréquences ne doivent pas être combinés directement. L'addition d'un tableau C1/m1 et d'un tableau C2/m2 définit une opinion collective dans laquelle le groupe des m1 juges a autant de poids que le groupe des m2 juges (il faudrait alors diviser par deux le résultat pour obtenir un tableau de fréquences).

- On peut attribuer à un juge k une importance ou un pouvoir particuliers en représentant son opinion non pas par un tableau individuel mais par un tableau collectif; par exemple:

a) Pondérations
Donner à un juge k une importance p_k .
C'est par définition représenter l'opinion du juge k par une opinion collective C. Cette opinion collective est l'opinion

de p_k juges ayant la même opinion que le juge k. On a $C = p_k C^k$ et $m = p_k$.
Le tableau collectif d'opinions C du juge k est défini par:

$c_{ij} = p_k$ si i "est en relation avec" j pour le juge k.
$c_{ij} = 0$ sinon

b) "Pouvoir"
Donner à un juge k le pouvoir de p_k juges.
C'est par définition représenter l'opinion du juge k par une opinion collective C. Cette opinion collective est l'opinion de p_k juges que le juge k représente; les p_k juges n'ont pas forcément la même opinion que le juge k, et peuvent très bien avoir chacun une opinion différente.
On a $C = \Sigma C^k$ et $m = p_k$.
Le tableau collectif C d'opinion du juge k est défini par:

c_{ij} = le nombre de juges que k représente ayant mis i en relation avec j.

Cette quantité est comprise entre 0 et p_k, mais elle peut prendre toutes les valeurs intermédiaires; ce qui n'est pas le cas pour une pondération.

4. PROPRIETES DES TABLEAUX D'OPINIONS

4.1. PROPRIETES DES TABLEAUX INDIVIDUELS

Définissons les propriétés suivantes valables pour un tableau de relation binaire C^ℓ ou C'^ℓ quelconques. Les résultats correspondants aux tableaux C'^ℓ seront obtenus en remplaçant c^ℓ_{ij} par $(c'^\ell_{ij} + 1)/2$.

4.1.1.Définition de propriétés

1) Transitivité
elle est définie par:

$c^\ell_{ij} = 1$ et $c^\ell_{jk} = 1$ impliquent $c^\ell_{ik} = 1$

Que l'on peut écrire linéairement :

$$c^\ell_{ij} + c^\ell_{jk} - c^\ell_{ik} \leq 1 \qquad \text{pour } i\neq j, j\neq k, i\neq k$$

2) Antisymétrie
Elle est définie par:

$c^\ell_{ij} = 1$ entraine $c^\ell_{ji} = 0$

qui se met sous la forme linéaire suivante:

$$c^{\ell}_{ij} + c^{\ell}_{ji} \leq 1 \qquad \text{pour } i \neq j$$

3) Symétrie
Elle est définie par:

$$c^{\ell}_{ij} = 1 \text{ entraine } c^{\ell}_{ji} = 1$$

qui se met sous la forme linéaire suivante:

$$c^{\ell}_{ij} - c^{\ell}_{ji} = 0 \qquad \text{pour } i \neq j$$

4) Totalité
Elle est définie par:

Pour tout couple (i,j) l'un au moins des deux cas $c^{\ell}_{ij} = 1$ ou $c^{\ell}_{ji} = 1$ est vrai.

Ce qui se met sous la forme linéaire suivante:

$$c^{\ell}_{ij} + c^{\ell}_{ji} \geq 1 \qquad \text{pour } i \neq j$$

5) Intermédiarité
Elle est définie par:

$$c^{\ell}_{ij} = 1 \text{ et } (c^{\ell}_{jk} = 0 \text{ et } c^{\ell}_{kj} = 0) \text{ et } c^{\ell}_{ks} = 1 \text{ impliquent } c^{\ell}_{is} = 1$$

Ce qui s'exprime par:
i"préféré" à j , j et k non comparables, k "préféré" à s entrainent que i est "préféré" à s. En notant P la préférence et I l'incomparabilité cette propriété est quelquefois donnée par certains auteurs sous la forme: PIP $\subset$ P.

Ce qui se met sous la forme linéaire suivante(Michaud):

$$c^{\ell}_{ij} - (c^{\ell}_{jk} + c^{\ell}_{kj}) + c^{\ell}_{ks} - c^{\ell}_{is} \leq 1 \qquad \text{pour } i \neq j \neq k \neq s$$

Cette inégalité est équivalente à:

$$c^{\ell}_{ij} - (c^{\ell}_{jk} + c^{\ell}_{kj}) + c^{\ell}_{ks} - 1 \leq c^{\ell}_{is}$$

L'expression de gauche ne peut dépasser la valeur 1. Elle n'atteint la valeur 1 que si l'on a simultanément i "préféré" à j, j et k non comparables et k "préféré" à s; d'où le résultat.

6) Cohérence
Elle est définie par:

$c^{\ell}_{ij} = 1$ et $c^{\ell}_{jk} = 1$ impliquent que les 2 expressions ($c^{\ell}_{is} = 0$ et $c^{\ell}_{si} = 0$) et ($c^{\ell}_{sk} = 0$ et $c^{\ell}_{ks} = 0$) ne doivent être simultanément vérifiées.

Ce qui s'exprime par i "préféré" à j et j "préféré" à k ,

ainsi que i "incomparable" à s et s "incomparable " à k est interdit. Avec les mêmes notations qu'en 5), certains auteurs expriment cette propriété par: $P^2 \cap I^2 = \emptyset$.

Ce qui se met sous la forme linéaire suivante (Michaud):

$$(c^{\ell}_{is} + c^{\ell}_{si}) + (c^{\ell}_{sk} + c^{\ell}_{ks}) \geq c^{\ell}_{ij} + c^{\ell}_{jk} - 1 \qquad \text{i,j,k,s étant tous différents}$$

La terme de droite ne peut dépasser la valeur 1 qui est atteinte lorsque i est "préféré" à j et j "préféré" à k. Dans ce cas le terme de gauche doit être supérieur ou égal à 1; et, i et s d'une part et k et s d'autre part ne peuvent être simultanément non comparables.

Les formules linéaires précédentes étant des définitions caractérisent complétement les propriétés associées. En combinant les formules correspondant à certaines des propriétés de définition précédentes, on en déduit de nouvelles formules. Attention dans ce cas les nouvelles formules obtenues ne caractérisent ni ne définissent la conjonction de ces différentes propriétés, elles n'en sont que la déduction.

4.1.2.Combinaisons de propriétés

Proprietés déduites de :

7) Transitivité, antisymétrie
En combinant les propriétés de transitivité et d'antisymétrie, on déduit l'expression linéaire suivante:

$$c^{\ell}_{ij} + c^{\ell}_{jk} + c^{\ell}_{ki} \leq 2 \qquad \text{pour i,j,k tous différents}$$

8) Transitivité, totalité
En combinant les propriétés de transitivité et de totalité on déduit l'expression linéaire suivante:

$$c^{\ell}_{ij} + c^{\ell}_{jk} + c^{\ell}_{ki} \geq 1 \qquad \text{pour i,j,k tous différents}$$

9) Antisymétrie, totalité
En combinant les propriétés d'antisymétrie et de totalité on déduit l'expression linéaire suivante:

$$c^{\ell}_{ij} + c^{\ell}_{ji} = 1 \qquad i \neq j$$

10) Transitivité, antisymétrie, totalité
Comme cette propriété contient les propriétés 7) et 8) on vérifie simultanément:

$$1 \leq c^{\ell}_{ij} + c^{\ell}_{jk} + c^{\ell}_{ki} \leq 2 \quad \text{pour } i,j,k \text{ différents}$$

En combinant les trois propriétés de transitivité d'antisymétrie et de totalité on obtient les inégalités doubles suivantes:

$$c^{\ell}_{ij} + c^{\ell}_{jk} - 1 \leq c^{\ell}_{ik} \leq c^{\ell}_{ij} + c^{\ell}_{jk}$$

Remarque:
Toutes les expressions, de 1) à 10), caractérisant soit la définition de propriétés (4.1.1) soit la combinaison de propriétés (4.1.2) sont des formules linéaires en termes des différents coefficients du tableau de comparaisons C^{ℓ} (ou du tableau C'^{ℓ} si l'on remplace c^{ℓ}_{ij} par $(c^{\ell}_{ij} + 1)/2$).

4.1.3.Définition de relations

Définissons quelques relations transitives classiques:

a)Relation de préordre:
C'est une relation, correspondant à un tableau C^{ℓ}, qui n'est que transitive.
Elle est définie par la formule donnée en 1).

b)Relation de préordre total:
C'est une relation qui est à la fois transitive et totale .
Elle est définie par les formules données en 1) et 4).

c)Relation d'ordre :
C'est une relation qui est à la fois transitive et antisymétrique.
Elle est définies par les formules données en 1) et 2)

d)Relation d'ordre total:
C'est une relation qui est à la fois transitive antisymétrique et totale.
Elle est définie par les formules données en 1), 2) et 4).

e)Relation d'équivalence:
C'est une relation qui est à la fois transitive et symétrique.
Elle est définie par les formules données en 1) et 3).

f)Relation d'équivalence totale:
C'est une relation qui est à la fois transitive symétrique et totale.
Elle est définie par les formules données en 1), 3) et 4).

Cette relation n'a aucun intérêt pratique et ne sera pas considérée par la suite; elle correspond à l'unique configuration où tous les éléments sont équivalents et appartiennent à une seule et même classe.

g)Relation d'ordre d'intervalles:
C'est une relation qui est à la fois transitive antisymétrique et qui possède la propriété d'intermédiarité.
Elle est définie par les formules données en 1), 2) et 5).

h)Relation de quasi-ordre (semiorder) :
C'est une relation qui est à la fois transitive antisymétrique et qui possède les propriétés d'intermédiarité et de cohérence.
Elle est définie par les formules données en 1), 2), 5) et 6).
On notera qu'un quasi-ordre est ici défini de façon antisymétrique, et correspond donc au terme anglais de semiorder, alors qu'il peut être aussi défini comme une relation totale, formalisation généralement utilisée dans (12).

Il existe également d'autres types de relations binaires classiques mais non transitives; citons par exemple:

i)Relation de tournoi:
C'est une relation qui est à la fois antisymétrique et totale.
Elle est définie par les formules données en 2) et 4) qui sont équivalentes à la formule donnée en 8).

j)Relation de match:
C'est tout simplement une relation totale.
Elle est définie par la formule 4).

Remarque:
Notons que toutes les relations binaires considérées ici sont définies exclusivement à partir des formules correspondant à la définition des propriétés présentées en (4-1-1). Comme nous l'avons vu ces formules sont toutes linéaires; il en résulte que ces relations binaires peuvent être définies en termes d'égalités ou d'inégalités linéaires, par rapport aux coefficients des tableaux de données C^k ou C'^k.

4.1.4.Obtention des tableaux C^k

A priori les tableaux d'opinions individuels C^k et C'^k peuvent n'avoir aucune autre propriété que celle de valoir 0 ou 1 pour c^k_{ij} et celle de valoir 1 ou -1 pour c'^k_{ij} .

Cependant dans la pratique l'obtention des tableaux C^k ou C'^k peut se faire directement ou par l'intermédiaire de tableaux de données annexes. Dans ce dernier cas les tableaux C^k ou C'^k auront souvent des propriétés particulières; donnons quelques exemples.

a) Cas de l'agrégation de préférences

1)Dans le cas le plus simple les données brutes se présentent sous forme de tableaux de classements (avec ou sans ex-aequo); le rang de l'objet i pour le juge k étant

représenté par r_{ik}. Le tableau C^k correspondant sera obtenu de la façon suivante:

$c^k_{ij} = 1$ si $r_{ik} \leq r_{jk}$

$c^k_{ij} = 0$ sinon

2)Si maintenant les données brutes se présentent sous la forme d'échelles ou de notes n_{ik} (note du juge k pour l'objet i), le tableau C^k associé pourra avoir l'une des formes suivantes selon les hypothèses:

$c^k_{ij} = 1$ si $n_{ik} \geq n_{jk}$

$c^k_{ij} = 0$ sinon

Dans ce cas le tableau C^k représentera un préordre total.

$c^k_{ij} = 1$ si $n_{ik} > n_{jk}$

$c^k_{ij} = 0$ sinon

Dans ce cas le tableau C^k représentera un ordre pouvant être partiel. Notons cependant que dans ce cas l'incomparabilité, correspondant à l'égalité des notes, est transitive.

$c^k_{ij} = 1$ si $n_{ik} > n_{jk} + s$
s étant un seuil positif ou nul

$c^k_{ij} = 0$ sinon

Dans ce cas le tableau représente un ordre en général partiel. Notons que dans ce cas l'incomparabilité n'est pas en général transitive, mais vérifie les propriétés d'intermédiarité et de cohérence 5) et 6) de (4-1-1), il s'agit donc d'un tableau représentant un quasi-ordre.

b) Cas de l'agrégation de similarités

1) Le cas le plus simple correspond à la situation où un juge représente un "critère" à plusieurs modalités. Le tableau de données brutes associé est alors un tableau de modalités, où r_{ik} représente la valeur de la modalité attribuée à l'objet i par le juge ou critère k. Le tableau C^k est défini de la façon suivante:

$c^k_{ij} = 1$ si $r_{ik} = r_{jk}$ c'est à dire i et j ont même modalité pour le juge k

$c^k_{ij} = 0$ sinon

Dans ce cas le tableau C^k représente une partition.

Dans le cas où le juge ou critère k représente une propriété, le tableau de données brutes associé est alors un tableau de présence-absence, où $r_{ik} = 1$ si l'objet i a la propriété k et $r_{jk} = 0$ si l'objet i n'a pas la propriété k. Dans ce cas le tableau C peut se définir de plusieurs façons:

$c^k_{ij} = 1$ si $r_{ik} . r_{jk} = 1$ c'est à dire si i et j ont tous les deux la propriété k
$c^k_{ij} = 0$ sinon

$c^k_{ij} = 1$ si $r_{ik} = r_{jk}$ c'est à dire si i et j ont ou n'ont pas simultanément la propriété k
$c^k_{ij} = 0$ sinon

Dans les deux cas le tableau C^k représente une partition.

2) Dans le cas où les données brutes sont représentées par un tableau de "proximités" P^k (représentant des distances, des dissimilarités, ou des mesures de proximités non nécessairement symétriques), le tableau C^k est défini de la façon suivante:

$c^k_{ij} = 1$ si $p^k_{ij} \leq s$, s étant un seuil donné
$c^k_{ij} = 0$ sinon

Le tableau C^k peut très bien ne pas représenter une partition, par exemple lorsque P^k n'est pas symétrique.

4.2. PROPRIETES DES TABLEAUX COLLECTIFS C ET C'

Comme nous l'avons vu précédemment les tableaux collectifs C et C' sont obtenus respectivement en sommant les tableaux individuels C^k et C'^k pour l'ensemble des m juges.

Au Chapitre 4-1 nous avons défini des relations et des propriétés individuelles associées aux tableaux C^k et C'^k, dans ce chapitre nous définirons cette fois des relations et des propriétés collectives associées aux tableaux C et C'
Pour ce faire nous utiliserons une règle de majorité " ma" qui nous permettra de définir une relation collective, à partir du tableau C (ou C').
La majorité ma est définie par ma = p m où p représente un pourcentage strictement positif; $0 < p \leq 1$. Sauf mention contraire p = 1/2, c'est la règle de la majorité de Condorcet. Mais toute autre valeur de ma est possible, en particulier si p = 1 on retrouve la règle de Pareto ou règle de l'unanimité.

Cette relation collective est alors définie de la façon suivante:

i "est en relation " avec j dans le relation collective , si et seulement si $c_{ij} \geq ma$; c'est à dire s'il y a au moins ma juges ayant mis i en relation avec j.

En général la relation collective dépend de la valeur ma de la majorité.

Remarque: Si m = 1 , cas d'un tableau individuel C^k, ma = p. Dans ce cas $c_{ij} \geq ma$ est indépendant de ma et équivalent à $c_{ij} = 1$.
On retrouve la définition des relations individuelles. Mais dans ce cas, ces relations individuelles sont indépendantes de la valeur de la majorité.

4.2.1.Définition de propriétés sur la relation collective
La définition se fera à partir des tableaux collectifs C et C'

1) Transitivité
Elle est définie par:

$c_{ij} \geq ma$ et $c_{jk} \geq ma$ impliquent $c_{ik} \geq ma$

2) Antisymétrie
Elle est définie par:

$c_{ij} \geq ma$ implique $c_{ji} \geq ma$

3) Symétrie
Elle est définie par:

$c_{ij} = c_{ji}$

4) Totalité
Elle est définie par:

Au moins l'une des inégalités $c_{ij} \geq ma$ ou $c_{ji} \geq ma$ est vérifiée.
Ce qui se traduit par $c_{ij} + c_{ji} \geq 2ma$

5) Intermédiarité
Elle est définie par:

$c_{ij} \geq ma$ et $c_{jk} < ma$ et $c_{kj} < ma$ et $c_{ks} \geq ma$ impliquent $c_{is} \geq ma$

6) Cohérence
Elle est définie par:

Les inégalités suivantes ne doivent pas être toutes

simultanément vérifiées:

$c_{ij} \geq ma$ et $c_{jk} \geq ma$ et $c_{is} < ma$ et $c_{si} < ma$ et $c_{sk} < ma$ et $c_{ks} < ma$

Remarque:
Certains auteurs considèrent d'autres types de propriétés dites "floues" . Ces propriétés sont associées à un tableau "flou" qui serait à rapprocher du tableau de fréquences C/m. Attention ces propriétés floues ne sont pas définies nécessairement à partir de la règle de majorité précédemment définie.
Par exemple (les propriétés seront définies relativement au tableau C):

7) Transitivité floue
Elle est définie par:

$$c_{ik} \geq \text{Min}(c_{ij} , c_{jk})$$

8) Transitivité stochastique
Elle est définie par:

$c_{ij} \geq m/2$ et $c_{jk} \geq m/2$ impliquent $c_{ik} \geq \text{Max} (c_{ij} , c_{jk})$

c'est à dire $c_{ik} \geq c_{ij}$ et $c_{ik} \geq c_{jk}$

9) Antisymétrie floue
Elle est définie par:

$c_{ij} = c_{ji}$ implique $c_{ij} = 0$ soit $c_{ij} \geq 0$ et $-c_{ij} \geq 0$

10) Antisymétrie floue parfaite
Elle est définie par:

$c_{ij} \geq c_{ji}$ implique $c_{ji} = 0$ soit $c_{ji} \geq 0$ et $-c_{ji} \geq 0$

etc..

4.2.2.Linéarisation de ces propriétés (Michaud)

On a le résultat suivant:
Toutes les propriétés de 1) à 10) peuvent se mettre sous la forme d'un système d'égalités ou d'inégalités linéaires contenant les entiers c_{ij} ainsi que des variables logiques x_{ij}.
Les conditions 3)de symétrie et 4) de totalité étant d'ailleurs initialement linéaires,ne nécessiteront pas une telle mise en forme.

Ceci résulte des conditions suivantes:

Par la suite nous considérerons des expressions

$a_i \geq 0$ i variant de 1 à q a_i étant un entier relatif
et $b_j \geq 0$ j variant de 1 à r
Ces expressions peuvent dépendre de variables quelconques et varier arbitrairement. La seule hypothèse étant qu'elles sont toutes bornées inférieurement et supérieurement. Nous appellerons par la suite t la plus grande de ces bornes. Par hypothèse on a:

$t \geq a_i \geq -t$ i variant de 1 à q
et $t \geq b_j \geq -t$ j variant de 1 à r

CONDITION 1
La vérification simultanée des q inégalités du système $a_i \geq 0$ i = 1...q est interdite.
Cette condition admet la représentation linéaire suivante:

$$(i) \quad (t+1)x_i \geq a_i + 1 \quad i = 1....q$$

$$(ii) \quad \Sigma\, x_i \leq q-1$$

A chaque inégalité $a_i \geq 0$ est associée une variable logique x_i valant 0 ou 1. Pour $a_i < 0$ x_i peut valoir 0 ou 1 tandis que $a_i \geq 0$ entraine $x_i = 1$ (relation (i)). (ii) exprime qu'il ne peut y avoir plus de q-1 x_i simultanément égaux à 1. D'où le résultat.
Le nouveau système contient le nombre d'inégalités et de variables initiales plus autant de variables logiques x_i que d'inégalités plus une nouvelle inégalité.

CONDITION 2
La vérification simultanée des q inégalités $a_i \geq 0$ pour i=1...q implique la vérification simultanée des r inégalités $b_j \geq 0$ pour j=1...r.
Cette condition admet la représentation linéaire suivante:

$$(i) \quad (t+1)x_i \geq a_i + 1 \qquad i=1...q$$

$$(ii) \quad b_j \geq (t+1)(\Sigma\, x_i - q) \quad j=1...r$$

Comme précédemment à chaque inégalité $a_i \geq 0$ est associée une variable logique x_i valant 0 ou 1 de telle sorte que $a_i \geq 0$ entraine $x_i=1$ (relations (i)). Lorsque les q inégalités $a_i \geq 0$ sont vérifiées $\Sigma\, x_i - q = 0$ et (ii) devient $b_j \geq 0$. Dans les autres cas l'expression $\Sigma\, x_i - q$ est strictement négative. Comme par définition $b_j \geq -(t+1)$ la relation est alors toujours vérifiée indépendamment de la valeur de b_j.
Le nouveau système contient le même nombre de variables et d'inégalités que le système initial plus q variables logiques x_i.

Remarque:

Les conditions 1 et 2 contiennent également les cas

d'égalités. Dans ce cas il suffit de représenter $a_i = 0$ par 2 inégalités $a_i \geq 0$ et $-a_i \geq 0$, de même pour b_j.

En résumé:
-les propriétés 3) symétrie et 4) totalité sont linéaires initialement

-La propriété 7) transitivité floue a une structure spéciale

-Les propriétés 1) transitivité , 2) antisymétrie ,5) intermédiarité, 8) transitivité stochastique, 9) antisymétrie floue et 10) antisymétrie floue parfaite correspondent à la Condition 2.

-La propriété 6) cohérence correspond à la Condition 1

Toutes les propriétés correspondant à la Conditions 1 ou 2 sont donc linéaires.
Pour appliquer les Conditions 1 ou 2 il suffira de remplacer les inégalités $c_{ij} \geq ma$ et $c_{ji} < ma$ respectivement par $c_{ij} \geq m'$ et $c_{ji} < m''$; m'étant le plus petit entier supérieur ou égal à ma et m" étant égal à m'-1.

Donnons deux exemples:

a) Transitivité
La propriété $c_{ij} \geq ma$ et $c_{jk} \geq ma$ impliquent $c_{ik} \geq ma$
se met sous la forme linéaire suivante:

$$(m+1)\, x_{ij} \geq (c_{ij} - m') + 1$$

$$(m+1)\, x_{jk} \geq (c_{jk} - m') + 1$$

$$c_{ik} - m' \geq (m+1)(x_{ij} + x_{jk} - 2)$$

Les variables logiques x_{ij} valant 0 ou 1

Les indices i,j,k étant tous différents.

b) Antisymétrie
La propriété: $c_{ij} \geq m'$ implique $c_{ji} < m''$
se met sous la forme linéaire suivante:

$$(m+1)\, x_{ij} \geq (c_{ij} - m') + 1$$

$$(m'' - c_{ji}) \geq (m+1)(x_{ij} - 1)$$

Pour terminer montrons directement que la propriété 7) transitivité floue peut se mettre sous forme linéaire.

La propriété: $c_{ik} \geq \text{Min}\,(c_{ij}, c_{jk})$

se met sous la forme linéaire suivante:

$$(i)\quad mx_{ijk} \geq c_{ij} - c_{jk} \geq -m(1 - x_{ijk})$$

$$(ii)\quad c_{ik} \geq c_{jk} - m\,(1 - x_{ijk})$$

$$(iii)\quad c_{ik} \geq c_{ij} - mx_{ijk}$$

$$x_{ijk} = 0 \text{ ou } 1$$

D'après (i) $c_{ij} > c_{jk}$ entraine $x_{ijk} = 1$ dans ce cas la relation (ii) se réduit à $c_{ik} \geq c_{jk}$ (cette relation est toujours vérifiée si $x_{ijk} = 0$).

De la même façon $c_{ij} < c_{jk}$ entraine $x_{ijk}=0$ et cette fois ci la relation (iii) se réduit à $c_{ik} \geq c_{ij}$.

Dans le cas où $c_{ij} = c_{jk}$ $x_{ijk} = 0$ ou 1 etc...

En définitive les propriétés sont donc toutes linéarisables.

4.2.3.Propriétés des tableaux C déduites de celles des tableaux C^k

Lorsque chacun des tableaux C^k vérifie l'une des propriétés définies en (4-1-1), les éléments des tableaux C et C' vérifient les inégalités ou égalités suivantes:

1) Transitivité pour chacun des C^k,
elle implique:

$$\text{et} \begin{matrix} c_{ij} + c_{jk} - m \leq c_{ik} & \text{(tableau C)} \\ c'_{ij} + c'_{jk} - m \leq c'_{ik} & \text{(tableau C')} \end{matrix} \qquad \text{pour i,j,k tous différents}$$

2) Antisymétrie pour chacun des C^k
elle implique:

$$\text{et} \begin{matrix} c_{ij} + c_{ji} \leq m \\ c'_{ij} + c'_{ji} \leq 0 \end{matrix} \qquad \text{pour } i \neq j$$

3) Symétrie pour chacun des C^k
elle implique :

$$\text{et} \begin{matrix} c_{ij} = c_{ji} \\ c'_{ij} = c'_{ji} \end{matrix} \qquad \text{pour } i \neq j$$

4) Totalité pour chacun des C^k
elle implique:

$$c_{ij} + c_{ji} \geq m$$

$c'_{ij} + c'_{ji} \geq 0$ pour $i \neq j$

5) Intermédiarité pour chacun des C^k
elle implique:

$$c_{ij} - (c_{jk} + c_{kj}) + c_{ks} - c_{is} \leq m$$

et i,j,k,s tous différents

$$c'_{ij} - (c'_{jk} + c'_{kj}) + c'_{ks} - c'_{is} \leq 3m$$

6) Cohérence pour chaque tableau C^k
elle implique :

$$-m \leq - c_{ij} - c_{jk} + (c_{is} + c_{si}) + (c_{sk} + c_{ks})$$

et i,j,k,s tous différents

$$-4m \leq - c'_{ij} - c'_{jk} + (c'_{is} + c'_{si}) + (c'_{sk} + c'_{ks})$$

7) Transitivité, antisymétrie pour chacun des C^k
elle implique:

$$\text{et} \begin{matrix} c_{ij} + c_{jk} + c_{ki} \leq 2m \\ c'_{ij} + c'_{jk} + c'_{ki} \leq m \end{matrix} \quad i,j,k \text{ tous différents}$$

8) Transitivité, totalité pour chaque tableau C^k
elle implique:

$$\text{et} \begin{matrix} c_{ij} + c_{jk} + c_{ki} \geq m \\ c'_{ij} + c'_{jk} + c'_{ki} \geq -m \end{matrix} \quad i,j,k \text{ tous différents}$$

9) Antisymétrie, totalité pour chaque tableau C^k
En combinant les propriétés d'antisymétrie et de totalité on déduit:

$$\text{et} \begin{matrix} c_{ij} + c_{ji} = m \\ c'_{ij} + c'_{ji} = 0 \end{matrix} \quad i,j,k \text{ tous différents}$$

10) Transitivité, antisymétrie, totalité pour chaque tableau C^k
Comme cette propriété contient les propriétés 7) et 8), on en déduit:

$$\text{et} \begin{matrix} m \leq c_{ij} + c_{jk} + c_{ki} \leq 2m \\ -m \leq c'_{ij} + c'_{jk} + c'_{ki} \leq m \end{matrix} \quad i,j,k \text{ tous différents}$$

En combinant les trois propriétés de transitivité, d'antisymétrie et de totalité on déduit les doubles

inégalités suivantes:

$$c_{ij} + c_{jk} - m \leq c_{ik} \leq c_{ij} + c_{jk}$$

et

$$c'_{ij} + c'_{jk} - m \leq c'_{ik} \leq c'_{ij} + c'_{jk} + m$$

5. DEFINITION DU PROBLEME ET PROPRIETES

On se propose dans cette nouvelle étape de trouver une relation collective transitive particulière "approximant au mieux" les m tableaux individuels C^k.
Cette relation transitive particulière est l'inconnue du problème et nous la noterons Y.

Dans tous les cas la relation binaire collective cherchée sera représentée, comme les tableaux C^k, par un tableau binaire Y, de terme général:

$y_{ij} = 1$ Si i "est en relation" avec j dans dans la relation collective
$y_{ij} = 0$ Sinon

L'expression "est en relation avec" dépend du contexte,et peut représenter l'une des 7 relations transitives définies en 4.1.3. Bien évidemment le tableau Y peut aussi représenter un tournoi ou un match; dans ce cas, la solution du problème est évidente.
Par exemple $y_{ij} = 1$ signifie i est classé avant j dans le cas où Y représente un ordre total tandis qu'il signifie i est équivalent à j si Y représente une relation d'équivalence.

Il nous faut maintenant définir ce que nous entendons par "approximer au mieux".
Nous définirons la distance $d(Y, C^k)$ entre un tableau C^k et la relation inconnue Y.
Nous définirons ensuite ce que nous appellerons la "distance" entre l'ensemble des tableaux collectifs C^k et la relation Y.

Le problème consistera alors à rechercher la relation collective Y située à distance minimale de l'ensemble des m relations C^k.

5.1. LES DISTANCES ASSOCIEES

5.1.1.Distances entre Y et C^k

La distance de Hölder entre la relation binaire collective cherchée Y et un tableau de relation C^k, est donnée par:

$$d_\ell^\ell(Y, C^k) = \sum_i \sum_j |y_{ij} - c^k_{ij}|^\ell$$

Cette distance a les propriétés suivantes:

1) Cette distance est linéaire en terme de Y .
La linéarité est due à la séparabilité de l'expression précédente en fonction de Y .
Considérons, en effet, une fonction séparable arbitraire de Y, elle peut s'écrire sous la forme:

$$f(Y) = \sum_i \sum_j f_{ij}(y_{ij})$$

Mais puisque y_{ij} est égal à 0 ou à 1, cette précédente expression est égale à:

$$f(Y) = \sum_i \sum_j f_{ij}(0) + \sum_i \sum_j (f_{ij}(1) - f_{ij}(0))y_{ij}$$

qui est bien une expression linéaire en Y . Ce résultat s'applique à la distance $d_\ell^\ell (C^k, Y)$, qui est une fonction séparable de Y; on obtient donc:

$$d_\ell^\ell(C^k,Y) = \sum \sum |c_{ij}^k|^\ell + \sum \sum(|1 - c_{ij}^k|^\ell - |c_{ij}^k|^\ell)y_{ij}$$

2) Toutes les distances de Hölder sont indépendantes de la caractéristique 1 de la métrique.
Puisque c_{ij}^k est égal à 0 ou 1, $|c_{ij}^k|^\ell$ et $|1 - c_{ij}^k|^\ell$ sont égaux respectivement à c_{ij}^k et $1 - c_{ij}^k$; et l'expression de $d_\ell^\ell (C^k ,Y)$ est égale à :

$$\sum_i \sum_j c_{ij}^k + \sum_i \sum_j (1- 2c_{ij}^k) y_{ij}$$

finalement l'expression de $d_\ell^\ell(C^k, Y)$ est indépendante de ℓ; elle est linéaire par rapport aux y_{ij} et est donnée par:

$$\sum_i \sum_j c_{ij}^k - \sum_i \sum_j c_{ij}^{\prime k} y_{ij}$$

Remarque:
Dans le langage de la théorie des ensembles, $d_1(C^k, Y)$ représente la distance de la différence symétrique entre deux ensembles. Les deux ensembles sont C^k et Y, leurs élements sont représentés par des 1 dans chaque tableau. La différence symétrique entre deux ensembles est égale au nombre d'éléments dans un des ensembles mais pas dans l'autre, c'est précisemment la valeur donnée par :

$$\sum_i \sum_j c_{ij}^k - \sum_i \sum_j c_{ij}^{\prime k} y_{ij}$$

5.1.2. "Distances" entre Y et tous les tableaux de relation C^k

a) Nous définirons la "distance" entre Y et tous les C^k par l'expression homogène suivante:

$$\sqrt[\ell]{\sum_k d_\ell^\ell (Y, C^k)}$$

C'est cette expression que la relation Y doit minimiser.

Or minimiser la précédente fonction est équivalent à minimiser (même solution) la fonction suivante:

$$\sum_{k=1}^{m} d_\ell^\ell (Y , C^k)$$

Cette fonction objectif ou économique, est indépendante de la caractéristique ℓ de la métrique de Hölder; elle est linéaire et égale à

$\sum_k d_1 (C^k, Y)$ pour $\ell = 1, 2....$

En particulier la distance quadratique ou euclidienne est en fait linéaire.

L'expression développée de cette fonction économique est donnée par:

$$\sum_i \sum_j \left(\sum_k c_{ij}^k\right) - \sum_i \sum_j \left(\sum_k c'^k_{ij} \right) y_{ij}$$

C'est à dire:

$$\sum_i \sum_j c_{ij} - \sum_i \sum_j c'_{ij} y_{ij}$$

Minimiser cette expression est équivalent à maximiser :

$$\sum_i \sum_j c'_{ij} y_{ij}$$

Interprétons cette dernière fonction économique. Supposons que i soit en relation avec j dans la relation collective Y. Dans ce cas, d'après la définition de C, nous avons c_{ij} accords; c'est à dire c_{ij} juges ont mis i en relation avec j. De la même façon, c'_{ij} représente le nombre d'accords moins le nombre de non-accords; c'est à dire le nombre de juges ayant mis i en relation avec j moins le nombre de juges n'ayant pas mis i en relation avec j. Si nous sommons toutes les valeurs $c'_{ij}\ y_{ij}$, pour tous les couples (i,j) et $i \neq j$, on obtient le nombre total d'accords moins le nombre total de non-accords entre l'ensemble des m relations C^k et la relation Y.

En résumé, le problème consistant à trouver la relation collective Y située à "distance" minimale de l'ensemble des relations C^k est équivalent à trouver la relation collective Y , maximisant le nombre total d'accords moins le nombre total de "non-accords" avec l'ensemble des opinions émises.

b) On peut également utiliser la "distance" entre Y et tous les tableaux C^k définie par:

$$\max_k \ d(Y, C^k)$$

Cette "distance" est en fait la limite quand la

caractéristique tend vers l'infini de la "distance" entre Y et les tableaux C^k définie précédemment à l'aide de la métrique de Holder.

Minimiser cette expression est équivalent à:

Minimiser z'

avec $z' \geq d(Y, C^k)$ pour k=1,...,m

Ce qui est aussi équivalent à:

Maximiser z

avec $z \leq \sum_i \sum_j c'_{ij} y_{ij}$ pour k=1,...,m

Il s'agit également d'un problème linéaire, cette fois en termes de z et des y_{ij}.

5.2. UNANIMITE ET LINEARITE
(Propriété paretienne)

1) Agrégation de tableaux individuels C^k transitifs par un tableau collectif Y transitif et total (préordre total).

On a le résultat suivant(Michaud):

Si j est en relation avec i pour l'ensemble des m juges et si i est en relation avec j pour un nombre de juges strictement inférieur à m, alors j est en relation avec i dans la solution collective optimale; c'est à dire le rang de j est inférieur ou égal au rang de i dans la solution optimale.

Appelons Y^1 la solution collective optimale. Dans ce cas le résultat précédent s'exprime par:

$c'_{ji} = m$ et $c'_{ij} < m$ entraine $y^1_{ji} = 1$

Démonstration:

Supposons que $c'_{ji} = m$ que $c'_{ij} < m$ mais que l'on ait $y^1_{ji} = 0$.
Comme Y^1 est total $y^1_{ij} = 1$ et i est classé strictement avant j dans le classement optimal Y^1.

Les n objets à classer se présentent comme suit dans le classement optimal Y^1:
Les objets strictement classés avant i (et avant j)
L'objet i
Les objets classés entre i et j (i et j exceptés)
L'objet j
Les objets classés strictement après j (et après i)

Appelons K l'ensemble des objets classés entre i et j (sans i et j). Cet ensemble se décompose en trois sous-ensembles disjoints définis de la façon suivante :
K_1 est l'ensemble des objets ex-aequo à i
K_2 est l'ensemble des objets classés strictement après i mais strictement avant j.
K_3 est l'ensemble des objets ex-aequo à j.

K_2 peut très bien ne pas contenir d'éléments.
De même pour K_1 et K_3; ce qui est le cas lorsque Y représente un ordre total.

La valeur de la fonction économique est donnée par:

$$F(C,Y^1) = Cte + c'_{ij} + \sum_{k\in K_1\cup K_2\cup K_3} c'_{ik} + \sum_{k\in K_1\cup K_2\cup K_3} c'_{kj} + \sum_{k\in K_1} c'_{ki} + \sum_{k\in K_3} c'_{jk}$$

Les termes contenant les indices i ou j sont indiqués, la partie constante contient tous les autres termes ou ni i ni j ne figurent.

Si l'on échange dans le préordre total Y^1 la position de i et de j on obtient un nouveau préordre total Y^2, où les ex-aequo de i deviennent les ex-aequo de j etc...
La valeur de la fonction économique correspondante est donnée par:

$$F(C,Y^2) = Cte + c'_{ji} + \sum_{k\in K_1\cup K_2\cup K_3} c'_{jk} + \sum_{k\in K_1\cup K_2\cup K_3} c'_{ki} + \sum_{k\in K_1} c'_{kj} + \sum_{k\in K_3} c'_{ik}$$

La constante est la même que précédemment. De plus les

expressions $\sum_{k\in K_1} c'_{ki}$, $\sum_{k\in K_3} c'_{jk}$, $\sum_{k\in K_1} c'_{kj}$ et $\sum_{k\in K_3} c'_{ik}$

apparaissent dans les deux fonctions économiques.

En retranchant $F(C,Y^1)$ de $F(C,Y^2)$ ces expresssions disparaissent et l'on obtient (en regroupant terme à terme):

$$F(C,Y^2)-F(C,Y^1)=c'_{ji} - c'_{ij} + \sum_{k\in K_1\cup K_2}(c'_{jk}-c'_{ik}) + \sum_{k\in K_1\cup K_2}(c'_{ki}- c'_{kj})$$

Comme par hypothèse chaque C'^k est transitif le tableau collectif C vérifie (4-2-2-1):

$$c'_{ji} + c'_{ik} - m \leq c'_{jk} \quad \text{d'une part}$$

et $c'_{kj} + c'_{ji} - m \leq c'_{ki}$ d'autre part.

Comme $c'_{ji} = m$ on a respectivement:

$$c'_{ik} \leq c'_{jk} \quad \text{et} \quad c'_{kj} \leq c'_{ki}$$

Dans l'expression de $F(C,Y^2)- F(C,Y^1)$ les termes entre parenthèses sont donc tous positifs ou nuls; comme d'autre part $c'_{ji} - c'_{ij} > 0$ par hypothèse, il en résulte que l'expression $F(C,Y2)-F(C,Y^1)$ est strictement positive et Y^1

n'est pas solution optimale, ce qui contredit l'hypothèse $y^1_{ji} = 0$. D'où le résultat.

2) Agrégation de tableaux individuels C^k transitifs par un tableau collectif Y antisymétrique transitif et total (ordre total)

Le résultat est identique au cas précédent ainsi que la démonstration. Dans ce cas K_1 et K_3 ne contiennent pas d'éléments.

3) Autres résultats (Michaud)

Les résultats précédents sont vrais également lorsque la distance entre Y et C^k est définie par:

$$d(C,Y) = \operatorname{Max}_k d(C^k,Y)$$

Dans ce cas la solution optimale Y est définie par:

$$d(C,Y^1) = \operatorname{Min} \operatorname{Max}_k d(C^k,Y)$$

Les résultats antérieurs sont vrais en prenant m = 1; c'est à dire lorsque le tableau collectif C représente l'un quelconque des tableaux individuels C^k.

Avec les mêmes notations que précédemment (Y^1 solution optimale) on a:

$$F(C^k,Y^2) - F(C^k,Y^1) \geq 0 \quad \text{si } c'^k_{ji} = 1 \text{ et } c'^k_{ij} = 1$$

$$\text{et } F(C^k,Y^2) - F(C^k,Y^1) > 0 \quad \text{si } c'^k_{ji} = 1 \text{ et } c'^k_{ij} = -1$$

(Le terme $c'^k_{ji} - c'^k_{ij}$ étant respectivement nul ou égal à 2 dans l'expression de $F(C^k,Y^2) - F(C^k,Y^1)$ définie précédemment en 1)).

Comme $d(C^k,Y) = \sum_i \sum_j c_{ij} - F(C^k,Y)$ on a :

$$d(C^k,Y^2) \leq d(C^k,Y^1) \quad \text{si } c'^k_{ji} = 1 \text{ et } c'^k_{ij} = 1$$

$$\text{et } d(C^k,Y^2) < d(C^k,Y^1) \quad \text{si } c'^k_{ji} = 1 \text{ et } c'^k_{ij} = -1$$

De plus $c'_{ji} = m$ et $c'_{ij} < m$ impliquent que $c'^k_{ji} = 1$ et $c'^k_{ij} = -1$ pour au moins un indice k. Pour cet indice l'inégalité stricte est vérifiée.

Ces résultats sont valables pour n'importe quel tableau C^k.

On a donc $\operatorname{Max}_k d(C^k,Y^2) < \operatorname{Max}_k d(C^k,Y^1)$

ce qui contredit l'optimalité de Y^1. D'où le résultat.

4) Agrégation de tableaux individuels C^k transitifs par un tableau collectif Y transitif et symétrique (équivalence).

On a le résultat suivant:
Si j est en relation avec i pour l'ensemble des m juges et si i est en relation avec j pour l'ensemble des m juges, alors j est en relation avec i dans le solution collective optimale; c'est à dire j et i sont dans la même classe dans la solution optimale.

Appelons encore Y^1 la solution optimale. Ce résultat s'exprime par:

$$c'_{ij} = m \text{ et } c'_{ji} = m \text{ impliquent } y^1_{ij} = 1$$

Démonstration:

Supposons que $c'_{ij} = m$ et $c'_{ji} = m$ mais que $y^1_{ij} = 0$. Comme Y^1 est symétrique $y^1_{ji} = 0$ c'est à dire i et j ne sont pas dans la même classe.

Appelons:
K_i les objets qui sont dans la même classe que i (i excepté)
K_j les objets qui sont dans la même classe que j (j excepté)
(Ces ensembles peuvent très bien être vides).

La valeur de la fonction économique est donnée par :

$$F(C,Y^1) = Cte + \sum_{k\in K_i} (c'_{ik} + c'_{ki})$$

Les termes contenant les indices i sont indiqués. La partie constante contient tous les termes où l'indice i ne figure pas.

Appelons Y^2 la solution collective dans laquelle i est mis dans la même classe que j, la valeur correspondante de la fonction économique est donnée par:

$$F(C,Y^2) = Cte + \sum_{k\in K_j} (c'_{ik} + c'_{ki}) + c'_{ij} + c'_{ji}$$

La constante étant la même que précédemment.

Posons:

$$S_i = \sum_{k\in K_i} (c'_{ik} + c'_{ki}) \quad (\text{Si } K_i = \emptyset,\ S_i = 0)$$

$$\text{et} \quad S_j = \sum_{k\in K_j} (c'_{ik} + c'_{ki}) \quad (\text{Si } K_j = \emptyset;\ S_j = 0)$$

a) $S_j \geq S_i$

Dans ce cas $F(C,Y2) - F(C,Y^1) = S_j - S_i + c'_{ij} + c'_{ji} \geq 2m$ et Y^1 n'est pas solution optimale.

b) $S_i > S_j$

Dans ce cas écrivons la fonction économique associée à Y^1 sous la forme suivante :

$$F(C,Y^1) = Cte + \sum_{k\in K_j} (c'_{jk} + c'_{kj})$$

Les termes contenant les indices j, cette fois, sont indiqués. La partie constante contient tous les termes où l'indice j ne figure pas

Appelons Y^3 la solution collective dans laquelle j est mis dans la même classe que i, la fonction économique associée est donnée par:

$$F(C,Y^3) = Cte + \sum_{k\in K_i}(c'_{jk} + c'_{kj}) + c'_{ij} + c'_{ji}$$

La constante étant la même que dans la formule précédente.

Comme par hypothèse chaque tableau C'^k est transitif le tableau collectif C' verifie :

$c'_{ij} + c'_{jk} - m \leq c'_{ik}$ d'une part et $c'_{ji} + c'_{ik} - m \leq c'_{jk}$ d'autre part

Comme par hypothèse $c'_{ij} = m$ et $c'_{ji} = m$, on a respectivement:

$c'_{jk} \leq c'_{ik}$ et $c'_{ik} \leq c'_{jk}$ d'où $c'_{ik} = c'_{jk}$

De même de:

$c'_{kj} + c'_{ji} - m \leq c'_{ki}$ et $c'_{ki} + c'_{ij} - m \leq c'_{kj}$ on en déduit

$$c'_{ki} = c'_{kj}$$

Il en résulte que :

$$\sum_{k\in K_j}(c'_{jk} + c'_{kj}) = \sum_{k\in K_j}(c'_{ik} + c'_{ki}) = S_j$$

$$\sum_{k\in K_i}(c'_{jk} + c'_{kj}) = \sum_{k\in K_i}(c'_{ik} + c'_{ki}) = S_i$$

On a donc $F(C,Y^3) - F(C,Y^1) = S_i - S_j + c'_{ij} + c'_{ji} \geq 2m$

puisque $S_i - S_j > 0$ par hypothèse.

Dans ce cas également Y^1 n'est pas non plus solution optimale. D'où le résultat.

En résumé Y^1 ne peut être solution optimale, ce qui contredit l'hypothèse $y'_{ij} = 0$; d'où le résultat annoncé.

Remarque:
Les résultats présentés ici ne font aucune hypothèse, excepté la transitivité, quant à la nature des tableaux de données individuels.

Si par contre les tableaux de données individuels sont de même nature que le tableau collectif (préordres pour le cas 1) (et 3)) , ordres pour le cas 2) (et 3)) et partitions pour le cas 4), on retrouve des résultats connus donnés par J.P.Barthélemy (2) , J.Feldman (5) , B.Monjardet (10) et S.Régnier (11).

5.3. REPRESENTATION DE Y SOUS FORME DE CONTRAINTES LINEAIRES

Nous avons vu précédemment que la relation transitive cherchée Y était définie par un tableau binaire de terme général y_{ij}. Mais ces valeurs y_{ij} ne sont pas arbitraires et suivant la signification donnée à Y (Relations de préordre, d'ordre, de quasi-ordre ou d'équivalence) les valeurs y_{ij} devront vérifier un certain nombre de propriétés (transitivité, antisymétrie, symétrie, totalité, intermédiarité, cohérence).

Ces propriétés ont été définies en (4-1-1) pour les tableaux C^k et s'appliquent bien entendu au tableau Y (substituer y_{ij} à c^k_{ij} dans les formules définissant les propriétés).
Comme nous l'avons vu précédemment chaque propriété correspond à un système d'égalités ou d'inégalités linéaires. Dans le cas du tableau Y les égalités ou inégalités seront linéaires par rapport aux y_{ij}.

Donnons quelques exemples:

a) Si Y est une relation d'ordre total, Y devra vérifier le système suivant

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \qquad \forall\ (i\neq j, j\neq k, i\neq k)$$

$$y_{ij} + y_{ji} = 1 \quad \text{pour } i\neq j$$

$$y_{ij} = 0 \text{ ou } 1$$

b) Si Y est une relation d'équivalence , Y devra vérifier le système suivant:

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \qquad \forall\ (i\neq j, j\neq k, i\neq k)$$

$$y_{ij} - y_{ji} = 0 \quad \text{pour } i\neq j$$

$$y_{ij} = 0 \text{ ou } 1$$

6. LES MODELES ASSOCIES

6.1. MODELES BIVALENTS

Notre problème, consistant à trouver une relation collective maximisant le nombre total d'accords moins le nombre total de non-accords, peut être représenté par un programme linéaire à variables bivalentes.
Ce programme est défini de la façon suivante:
Les variables 0-1 sont les valeurs y_{ij} de la relation Y.
La fonction économique à maximiser représente le nombre total d'accords moins le nombre total de non-accords.
Les contraintes représentent les différentes propriétés que doivent vérifier la relation Y. Il s'agit de la transitivité et selon le cas des propriétés d'antisymétrie, de symétrie , de totalité, d'intermédiarité ou de cohérence.
La fonction économique $\Sigma\Sigma\, c'_{ij}\, y_{ij}$ est linéaire par rapport aux variables y_{ij} ainsi que les contraintes représentant les différentes propriétés de la relation Y. Il s'agit donc bien d'un programme linéaire.

En fonction de la relation Y recherchée on obtient les modèles suivants:

a) Modèle pour une relation d'ordre total:

$$\text{Max} \sum_i \sum_{j\neq i} c'_{ij}\, y_{ij}$$

$$y_{ij} + y_{ji} = 1 \qquad i\neq j$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \qquad i\neq j, j\neq k, i\neq k$$

$$y_{ij} = 0 \text{ ou } 1$$

Ce modèle a n(n-1) relations de totalité et antisymétrie, et n(n-1)(n-2) relations de transitivité.
Ce modèle peut être réduit; division par 2 du nombre de variables et approximativement par 3 du nombre de contraintes. (Voir (7) p 139)

Voici quelques propriétés de ce modèle:

1) Ce problème est équivalent dans le cas où Y est un ordre total à la maximisation du nombre d'accords. En effet $\Sigma\Sigma\, c'_{ij}\, y_{ij}$ est égal à $2\ \Sigma\Sigma\, c_{ij}\, y_{ij} - m\ \Sigma\Sigma\, y_{ij}$;et le terme $m\ \Sigma\Sigma\, y_{ij} = mn(n-1)/2$ puisque Y est un ordre total.
Ce problème est équivalent également à la minimisation du nombre de désaccords ($\Sigma\Sigma\, c_{ji}\, y_{ij}$), soit enfin à la maximisation du nombre d'accords moins le nombre de désaccords ($\sum_i \sum_{j\neq i} (c_{ij} - c_{ji})y_{ij}$). (Voir (7) p 137 ,où la

notion de désaccord est définie).

2) Chacune de ces fonctions économiques peut être divisée par la quantité ($\Sigma\Sigma\, c_{ij}$); ce qui donne des fonctions économiques normalisées équivalentes à la fonction économique originale; on retrouve ici les notions de "coefficient d'adhésion" introduit par E.Jacquet-Lagrèze, de "taux" d'accords ou de désaccords etc... (voir (7) p 178).

3) Dans le cas où les tableaux C^k ne représentent pas des ordres , même le cas d'un unique tableau de comparaisons est intéressant, c'est le problème classique de la théorie des graphes où l'on recherche un ordre à distance minimale d'une relation binaire, d'un tournoi par exemple.

b) Modèle dans le cas d'une relation d'équivalence

$$\text{Max} \quad \sum_{ij \neq i} \sum c'_{ij}\, y_{ij}$$

$$y_{ij} - y_{ji} = 0 \qquad i \neq j$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \qquad i \neq j, j \neq k, k \neq i$$

$$y_{ij} = 0 \text{ ou } 1$$

Ce modèle a $n(n-1)$ relations de symétrie et $n(n-1)(n-2)$ relations de transitivité. Comme dans le cas précédent le modèle peut être réduit, division par 2 du nombre des variables et division par 2 du nombre des contraintes, (voir (8) p 45).

Indiquons succinctement quelques propriétés de ce modèle :

1) Contrairement au cas où Y est un ordre total, ce problème n'est pas équivalent à la maximisation du nombre total d'accords

$$\Sigma\, \Sigma\, c_{ij}\, y_{ij}$$

ou à la minimisation du nombre total de non-accords

$$\Sigma\, \Sigma\, (m - c_{ij})\, y_{ij}$$

bien que la fonction économique $\Sigma\, \Sigma\, c'_{ij} \cdot y_{ij}$ soit par définition la différence de ces deux quantités.

Remarquons que lorsque Y est une partition , le problème de la maximisation du nombre d'accords a une solution évidente.

Il s'agit de la partition "grossière" où tous les éléments sont équivalents. Il en est de même du problème de la minimisation du nombre de non-accords qui lui aussi a une solution évidente. Il s'agit cette fois de la partition "triviale" où chaque élément forme une classe distincte.

Le critère de la partition centrale $c'_{ij} = c_{ij} + (c_{ij} - m)$ correspond à une fois le critère d'obtention de la partition grossière plus une fois le critère d'obtention de la partition triviale.
Notons (1,1) les coefficients de compromis linéaire pour la partition centrale . Par extension , on peut définir des partitions (a,b) centrales avec $c'_{ij} = a\, c_{ij} + b\,(c_{ij} - m)$ pour $a + b = 2$.

2) Les tableaux de similarités C^k peuvent représenter des partitions, dans ce cas particulier le problème est connu sous le nom du problème de la "partition centrale" posé par S. Régnier (voir (11)).

3) Lorsque les tableaux de similarités C^k ne représentent pas des partitions , même le cas d'un simple tableau de similarités est intéressant.C'est le problème de théorie des graphes où une relation binaire (le tableau C^k) est approximée par une relation d'équivalence (la relation Y).
Un cas particulier de ce problème est le problème de Zahn, où une relation symétrique est approximée par une relation d'équivalence.

c) Modèle pour une relation de quasi-ordre

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij}\, y_{ij}$$

$$y_{ij} + y_{ji} \leq 1 \qquad i \neq j$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \qquad i,j,k \text{ différents}$$

$$y_{ij} - (y_{jk} + y_{kj}) + y_{k\ell} - y_{i\ell} \leq 1 \qquad i,j,k,\ell \text{ différents}$$

$$y_{ij} + y_{jk} - (y_{i\ell} + y_{\ell i}) - (y_{ik} + y_{ki}) \leq 1 \qquad i,j,k,\ell \text{ différents}$$

Ce modèle a $n(n-1)$ variables y_{ij} et un nombre de contraintes de l'ordre de n^4 .

d) Autres modèles

De la même façon on pourra tout aussi bien définir les modèles où la relation collective Y représente un préordr

partiel, un préordre total, un ordre partiel, un ordre d'intervalles.

Contrairement aux cas de l'ordre total ou de l'équivalence , le nombre de variables ou de contraintes ne peut être réduit pour les autres modèles. Le nombre des variables reste égal à n(n-1)
Le nombre de contraintes est de l'ordre de n^3 (du aux relations de transitivité) sauf pour l'ordre d'intervalles où le nombre de contraintes est de l'ordre de n^4 (du aux relations d'intermédiarités).

6.2. PROGRAMMES LINEAIRES ASSOCIES ET RESOLUTION

Si nous remplaçons les conditions $y_{ij} = 0$ ou 1, par des conditions de variation continue $0 \leq y_{ij} \leq 1 \quad \forall \; i \neq j$ nous obtenons, à la place des modèles bivalents précédents, des programmes linéaires continus.
Ces programmes linéaires continus peuvent être résolus par programmation linéaire , par la méthode du simplexe par exemple.

Cependant, même la résolution du problème continu, sous sa forme initiale, parait difficile.
En effet la taille des bases, représentées par le nombre de contraintes, est de l'ordre de n^3 ou n^4 selon les cas; ce qui exclut toute possibilité pratique de résolution.
Par contre, la résolution du problème dual associé ,(voir (7)), est possible; dans ce cas la taille des bases étant égale à n(n-1) le nombre de variables y_{ij} du problème primal, il devient alors possible de résoudre des problèmes pour n relativement grand.

Disons simplement que d'un point de vue théorique , la résolution du problème dual donne la solution du problème primal continu.
Par contre le modèle bivalent et le modèle continu ne sont pas équivalents (voir (7)) et la possibilité d'obtenir une solution fractionnaire pour le programme linéaire associé existe. Mais ce phénomène est assez rare et ne semble pas se produire dans les problèmes pratiques.

Actuellement nous avons résolu par cette approche plus de 140 problèmes d'agrégation par un classement ou par une partition. Dans chaque cas, pour le problème continu, nous avons obtenu une solution optimale bivalente, donc solution optimale du problème bivalent.On trouvera quelques uns de ces exemples détaillés dans (1),(7),(9).

6.3. AUTRES MODELES

Lorsque l'on combine des tableaux individuels C^k on obtient un tableau collectif C.
Mais ce tableau ne conserve pas forcément les propriétés initiales que pouvaient avoir les tableaux C^k ; par exemple l'antisymétrie ou la transitivité ne sont pas nécessairement additives (effet Condorcet).

Il peut sembler intéressant d'approximer ce tableau collectif C par un tableau collectif inconnu Z. Ce tableau Z pouvant par exemple posséder les propriétés de transitivité, d'antisymétrie et de totalité collectives (définies en (4.2.1))(ordre total collectif). D'une façon générale on peut approximer le tableau collectif C par un tableau collectif Z vérifiant certaines des propriétés définies en (4.2.1).
Toutes ces propriétés qui représentent des contraintes sur le tableau Z sont des contraintes linéaires comme nous l'avons vu en (4.2.2). De même la fonction économique à minimiser est linéaire si elle est définie par:

$$\text{Min} \sum_i \sum_{j \neq i} |c_{ij} - z_{ij}|$$

qui se met sous la forme linéaire suivante:

$$\text{Min} \sum_i \sum_j u^+_{ij} + u^-_{ij}$$

avec

$$c_{ij} - z_{ij} = u^+_{ij} - u^-_{ij}$$

$$u_{ij} \geq 0$$

$$u_{ij} \geq 0$$

Les z_{ij} étant entiers.

La définition de ces problèmes sous forme de programmation linéaire en nombres entiers ne semble pas avoir été faite même dans le cas de la recherche d'un tableau collectif Z possédant la propriété de transitivité floue 7) donnée en (4.2.1). Par contre la résolution approchée de ce dernier problème est classique; elle est obtenue en recherchant la fermeture transitive de la relation collective C. Ce problème d'approximation se rencontre souvent en classification automatique sous le nom de recherche de l'ultramétrique sous-dominante lorsque le tableau C représente cette fois un tableau de dissimilarités.

7. REFERENCES

[1] AYACHE G.,MARCOTORCHINO J.F., MICHAUD P. "Démocratie Française et la Presse, Techniques d'Analyse de la

Réception d'un Message", Revue Française de Communication, No 1, 30, (1978).

[2] BARTHELEMY J.P., "sur les Eloignements Symétriques et le Principe de Pareto", Mathématiques et Sciences Humaines, No 56, 97, (1976).

[3] BARTHELEMY J.P., MONJARDET B., "Ajustement et Résumé de Données Relationnelles: Les Relations Centrales", (Article présenté aux 2ème Journées Internationales d' Analyse des Données de l'IRIA, Versailles, (1979)).

[4] DIDAY E., Nouvelles Méthodes et Nouveaux Concepts en Classification Automatique et Reconnaissance des Formes, Thèse de Doctorat d'Etat, Université de Paris VI, (1972).

[5] FELDMAN J., "Pôles,Intermédiaires et Centres dans un Groupe d'Opinions", Mathématiques et Sciences Humaines, No 43, 39, (1973).

[6] JACQUET-LAGREZE E., La Modélisation des Préférences; Préordres, Quasi-Ordres et Relations Floues , Thèse de 3ème Cycle, Université Paris V, (1975).

[7] MARCOTORCHINO J.F., MICHAUD P., Optimisation en Analyse Ordinale des Données , Masson, Paris (1979).

[8] MICHAUD P., MARCOTORCHINO J.F., "Optimization in Ordinal Data Analysis", (article présenté pour une partie au Symposium International d'Optimisation Discrète, Vancouver (1977),) et disponible dans Rapport Technique No1,IBM France Scientific Centre, (1978).

[9] MICHAUD P., MARCOTORCHINO J.F., "Optimisation en Analyse des Données Relationnelles", (Article présenté aux 2ème Journées Internationales d' Analyse des Données de l'IRIA, Versailles (1979)).

[10] MONJARDET B., "Tournois et Ordres Médians pour une Opinion", Mathématiques et Sciences Humaines, No 43, 55, (1973).

[11] REGNIER S., "Sur quelques Aspects Mathématiques des Problèmes de Classification Automatique", I.C.C. Bulletin No 4, 175, (1965).

[12] "Modélisation des préférences et quasi-ordres", Math. Sci. hum., n°62 et 63 (1978).

P3

OPTIMIZATION IN OPINION'S AGGREGATION

Pierre MICHAUD Francois MARCOTORCHINO
Centre Scientifique IBM-FRANCE, Paris

The aim of this presentation is to show that a certain number of aggregation problems can be represented by optimization models and solved efficiently by optimization techniques.

Key words: Linear Programming, Binary relations

1. INTRODUCTION

In this communication we present an optimization approach to solve the general problem of the aggregation of m binary relations, defined on a set of n objects, by a particular binary relation.

The particular binary relation considered will be whether various kinds of orders, this is the preference aggregation problem, or an equivalence relation, this is the similarity aggregation problem.

Moreover to define completly this problem, we must chose an aggregation rule allowing to define the "collective" relation from the m "individual" relations. This aggregation rule will be the "Condorcet's majority rule under constraints".

The problem is then: the individual opinions as well as the aggregation rule being given, find a collective opinion which is the unknown of the problem.

With a wide range of applicability and a "nice" aggregation rule this aggregation model seems attractive for a lot of practical applications. However, owing to the complexity of the computation of the collective opinion, this approach seemed unitilizable to solve real life problems until recently.

In this article, we shall show that effective computation based on optimization is possible for this aggregation model.

2. THE AGGREGATION MODEL

2.1 The individual opinions

An individual opinion can be interpretated as the opinion of a "judge". It is assumed that the opinion of a judge k, on the n objects, is defined in terms of pairwise comparisons; the n objects are compared two by two. By convention, an object will never be compared with itself, the number of comparisons is then n(n-1).

For every judge the result of the comparisons will be represented by a comparison or relational tableau C^k with:

c^k_{ij} = 1 if the object i is in relation with j for the judge k.

c^k_{ij} = 0 otherwise.

There is no other assumption on the opinion C^k of a judge k; it can be an arbitrary binary relation.
However, according to the context, "is in relation with" can mean "is preferred to" as in the case of preference aggregation or "is similar to" as in the case of similarity aggregation.

2.2 The collective opinion

As for the individual opinions, a collective opinion will also be represented by a comparison or relational tableau Y with:

y_{ij} = 1 if i is relation with j in the collective opinion

y_{ij} = 0 otherwise

And again, by convention, an object will never be compared with itself.
But, unlike an individual opinion, a collective opinion will not be arbitrary. It will satisfy a certain number of properties according to the type of aggregation considered.
Here is the list and the definition of properties that we shall consider for the collective opinion Y.

- Transitivity:
If y_{ij} = 1 and y_{jk} =1 then y_{ik} =1 for i,j,k all different.

- Symmetry:
y_{ij} = y_{ji} for i≠j

- Asymmetry:
At most y_{ij} = 1 or y_{ji} = 1 i≠j

- Completness:
At least y_{ij} = 1 or y_{ji} = 1 i≠j

- Intermediarity:
If y_{ij} = 1, y_{jk} = 0 and y_{ks} = 1 then y_{is} = 1

- Coherence:
If y_{ij} =1 and y_{jk} =1 then we cannot have at the same time (y_{is} =0 and y_{si} =0) and (y_{ks} =0 and y_{sk} =0).

From the above definition we can now define seven types of aggregation problems corresponding to seven types of collective opinion Y.

- (Partial) preorder:
It is just a transitive relation

- Complete preorder:
It is a transitive and complete relation

- (Partial) order:
It is a transitive and asymmetric relation

- Complete order:
It is a transitive, asymmetric and complete relation

- "Interval order":
It is a partial order satisfaying the intermediarity condition.

- Semi-order:
It is an interval order satisfaying the coherence condition.

- Partition or equivalence relation:
It is a transitive and symmetric relation.

The first six types, where the collective opinion Y represents various kinds of orders, correspond to preference aggregation problems. The most used and the most natural type, in practice, is the case where Y represents a complete order (ranking whithout ties).
The last type, where the collective opinion Y represents an equivalence relation, corresponds to the similarity aggregation problem.

2.3 The aggregation rule

This rule, the Condorcet's majority rule under constraints, consists of two parts:

1) An objective function, of the collective opinion Y, to maximize.
This is the aggregation criterion. This aggregation criterion will be the same for all the aggregation problems considered.

2) A set of constraints on the collective opinion Y.
This set of constraints will depend upon the aggregation problem considered.

Let us now define this rule :
If we add the C^k individual tableaux we obtain an opinion tableau C defined by:

$$C = \sum_k C^k$$

with

c_{ij} = The number of judges putting "i in relation with j"

To simplify the notations let us consider a C'^k tableau defined by:

$c'^k_{ij} = 2\, c^k_{ij} - 1$

$c'^k_{ij} = 1$ if i is in relation with j for the judge k

$c'^k_{ij} = -1$ otherwise

Similarly we have $C' = \sum_k C'^k$ with:

c'_{ij} = the number of judges putting i in relation with j minus the number of judges not putting i in relation with j.

$c'_{ij} > 0$ means that i is in relation with j for a (strict) majority of judges; $c'_{ij} < 0$ means that i is not in relation with j for a majority of judges.

1) The aggregation criterion is the following:

Maximize the objective function $\sum_i \sum_j c'_{ij}\ y_{ij}$ (2.1)

The maximization of this objective function (2.1) when Y is just an arbitrary binary relation, has a solution given by:

$y_{ij} = 1$ if $c'_{ij} > 0$

$y_{ij} = 0$ if $c'_{ij} < 0$

If $c'_{ij} = 0$ there is a tie and $y_{ij} = 0$ or 1

This is the Condorcet's majority rule.
This objective function has also an interpretation in term of distances between the collective opinion Y and the individual opinions C^k.
Let us define the "symmetric difference" distance between the Y tableau and the C^k tableaux by:

$$d(C^k, Y) = \sum_{ij} \sum |\ c^k_{ij} - y_{ij}\ |$$

This expression is equal to $\sum_i \sum_j c^k_{ij} - \sum_i \sum_j c'^k_{ij}\ y_{ij}$ (see [9],and the "distance" between Y and all the C^k tableaux is defined by:

$\sum_k d(C^k, Y)$ and this expression is equal to

$\sum_{ij} \sum c_{ij} - \sum_i \sum_j c'_{ij}\ y_{ij}$ (see [9]) (2.2)

Since this expression (2.2) is equal to minus the objective

function (2.1) plus a constant the minimization (for arbitrary constraints) of the expression (2.2) and the maximization of (2.1) are equivalent (they have the same solutions).

Then the criterion of minimizing this "distance" is nothing but the Condorcet's majority rule.
The unconstrained maximization of this criterion does not constitute an aggregation rule since there is no reason in general that the optimal unconstrained solution Y satisfies the constraints of the particular type of opinion under consideration.

For instance in the case of aggregation of total orders by a total order there is the possibility of $c'_{ij} > 0$ $c'_{jk} > 0$ and $c'_{ki} > 0$. Thus the optimal unconstrained solution satisfies $y_{ij} = 1$ $y_{jk} = 1$ and $y_{ki} = 1$ and therefore cannot represent a total order; this is the well known "Condorcet effect" or "Paradox of voting".

2) Definition of the set of constraints on the collective opinion Y: This set of constraints represents the various conditions satisfied by the particular type of collective opinions considered.
For instance the constraints on Y can be the ones of a total order or of a total preorder in a preference aggregation problem, or the ones of an equivalence relation in a similarity aggregation problem. If Y must be a total order the constraints will represent the conditions of transitivity, completness and asymmetry.

Finally the aggregation rule, the Condorcet's majority rule under constraints, is defined by:

Maximize $\Sigma \Sigma c'_{ij} y_{ij}$ for Y satisfaying the constraints of the aggregation problem.
Let us notice that Y is in general solution of a combinatorial problem and is therefore usually extremely difficult to compute. For instance if the collective opinion Y is a complete order, the number of feasible solutions Y is n!; this number has more than 100 digits for n=68.
If the collective opinion represents an equivalence relation, the number of feasible solutions is equal to B(n), the Bell's numbers; and $B(n) > \exp(n)$ for $n>7$.

For these reasons practical exact solving methods did not exist until now. Explicit enumeration methods are limited to problems with a very small number of objects 12, 10 or less, according to the type of problem. No other exact method in general has been proposed excepted in the case where Y represents a total order.
For this particular problem exact implicit enumeration

methods have been proposed, some of them are mentionned in [6], [9], they are limited in size to about 20 objects.

We now present an approach allowing to solve exactly much bigger problems.

3. THE PROPOSED LINEAR MODEL

The general aggregation model defined previously is an optimization model with a linear objective function and binary variables.
If it is possible to express also the constraints linearly then we obtain a linear program with binary variables. it is just what we have done in [6],[9] and [11], where we have shown that the various constraints of 2. can be expressed in terms of linear inequalities or equalities as follows:

- Transitivity: $y_{ij} + y_{jk} - y_{ik} \leqslant 1$
- Symmetry : $y_{ij} - y_{ji} = 0$
- Asymmetry: $y_{ij} + y_{ji} \leqslant 1$
- Completness: $y_{ij} + y_{ji} \geqslant 1$
- Intermediarity: $y_{ij} - y_{jk} + y_{ks} - y_{is} \leqslant 1$
- Coherence:
  $y_{ij} + y_{jk} - (y_{ks} + y_{sk}) - (y_{is} + y_{si}) \leqslant 1$

Thus all the six preference aggregation problems and the similarity aggregation problem defined in Chapter 2., can be represented by linear programs with binary variables see [6], [9] and [11].

As far as we know, this general approach has been proposed for the first time by P.Michaud and F.Marcotorchino in [6], [9] and [11].
In the particular case where all the opinions individual and collective are complete orders, a linear model, but with a different objective function, was first proposed prior to this general linear model by J. de Cani in [3] and later abandonned by the same author in [4].
All the proposed models are binary linear programs and not continuous linear programs. These binary programs are not, in general, equivalent to the linear programs where the constraints $y_{ij} = 0$ or 1 are replaced by $0 \leqslant y_{ij} \leqslant 1$. For

instance in the particular case of aggregation by a complete order this equivalence was falsely asserted by J.Bowman in [2] (it was also implicitely assumed by de Cani in [3]). This assertion has been contradicted by counterexemples given by A.Hoffman, E.Johnson and ourselves in [9].

Here are for instance two models: the preference aggregation model where the collective opinion Y represents a complete order and the similarity aggregation model where the collective opinion Y represents an equivalence relation (partition).

Model for a complete order relation Y

$$\begin{aligned} &\text{Max} \sum_i \sum_j c'_{ij}\, y_{ij} \\ &y_{ij} + y_{ji} = 1 \qquad i \neq j \qquad (3.1) \\ &y_{ij} + y_{jk} - y_{ik} \leqslant 1 \qquad i \neq j,\ j \neq k,\ i \neq k \\ &y_{ij} = 0 \text{ or } 1 \end{aligned}$$

Model for an equivalence relation Y

$$\begin{aligned} &\text{Max} \sum_i \sum_j c'_{ij}\, y_{ij} \\ &y_{ij} - y_{ji} = 0 \qquad i \neq j \qquad (3.2) \\ &y_{ij} + y_{jk} - y_{ik} \leqslant 1 \qquad i \neq j,\ j \neq k,\ i \neq k \\ &y_{ij} = 0 \text{ or } 1 \end{aligned}$$

4. THE PROPOSED SOLVING METHOD

With binary variables and a very large size, the binary linear programs (3.1) and (3.2) are problems a priori excessively complicated. However we have been able to obtain optimal solutions of these problems in a reasonnable amount of time even for sizes of n relatively important.

The proposed method is a cutting plane method. This method consists of a sequence of continuous linear programs where at each step a constraint (cut) is added to exclude the fractionnal solutions.

The initial L.P. is the binary linear program (3.1) or (3.2) where the binary constraints $y_{ij} = 0$ or 1 are replaced by $0 \leqslant y_{ij} \leqslant 1$. If the cuts are chosen adequately this exact method is finite, but it can be extremely long.

The number of cuts and the number of iterations of the L.P's can be very large. Contrary to what happens usually, it is not the number of continuous linear programs to solve which is the main difficulty for this problem (most of the time there is only one L.P. to solve) but the resolution of just one of these continuous linear programs.
Indeed the resolution of the linear programs (3.1) or (3.2) where $y_{ij}=0$ or 1 is replaced by $0 \leqslant y_{ij} \leqslant 1$ is a priori impossible if one tries to solve directly these programs, since the program is automatically turned into standard form by adding slack variables to use the Simplex method. The number of constraints being of order n-cubed, the size of the basic matrices is then $O(n^3)XO(n^3)$ and the handling and the inversion of such matrices is not possible even for n small, making the computation impossible.
For these reason the proposed method handles the linear program in the inequality form instead of standard form . The size of the bases is then $O(n^2)$ instead of $O(n^3)$ making possible the resolution of problems of much bigger size. To speed up the algorithm , we have also defined a good starting solution. The phase I is then by-passed and a phase II is started directly.
All this and other features of this method are detailed in [9].

Practically, for the computation we have used the I.B.M. program product MPSX/370 whose advantages such as extended control language ECL or degeneracy handling are also detailed in [9]. This has enable us to solve exactly problems of relatively large size, n=80 until now.
However, and this is a prominent point, for all the practical examples that we have considered (more than 200 preference and similarity problems at this time) the optimal solution of the first linear program (i.e whithout cuts) was always with binary values and therefore optimal solution of the binary linear program.
This fact of having to solve, most of the time, just one continuous linear program is a very outstanding empirical result, which justifies clearly by itself the optimization approach that we have chosen.

Computational results, with MPSX/370 on a system IBM/168, can be found in the following table: (solving times including Generation, Solving and Editing).

TYPE of PB	SIZE	NATURE	TIMES(Exact.Method)
Archaeology	21	P	7 s
Sociology	20	P	13 s
Medicine	23	P	15 s
Linguistics	25	R	7 s
Linguistics	28	R	11 s
Medicine	29	P	22 s
Linguistics	29	R	12 s

Esthetics	36	R	22 s
Psychology	35	P	36 s
Sociology	40	R	31 s
Compar.Tests	40	P	297 s
Linguistics	45	R	60 s
Linguistics	45	R	58 s
Voting	48	R	65 s
Voting	54	R	117 s
Politics	54	P	220 s
Sociology	72	R	157 s
Linguistics	77	R	331 s

(The uppercases "P" and "R" mean that the result is either a "partition" or a "ranking")

5. EXAMPLE OF APPLICATION

In this example the individual opinion tableaux are reduced to just one non transitive opinion tableau and the collective opinion tableau Y is a complete order.

This is a small example, adapted from a book of Jane Van Lawick Goodall "In the Shadow of Man". This book deals with the social behavior of Chimpanzees of the Gombe Stream National Park in Kenya. The author studied a group of 20 chimpanzees during a long period.

The attached list indicates the names given by Mrs Goodall to the 20 chimpanzees studied:

1 :	David	(M)	11 :	J.B.	(M)
2 :	Goliath	(M)	12 :	Huxley	(M)
3 :	Mc Gregor	(M)	13 :	Leakey	(M)
4 :	Flo	(F)	14 :	Hugh	(M)
5 :	Fifi	(F)	15 :	Rodolph	(M)
6 :	Figan	(A)	16 :	Humphrey	(M)
7 :	William	(M)	17 :	Evered	(A)
8 :	Olly	(F)	18 :	Worzle	(M)
9 :	Flint	(A)	19 :	Faben	(A)
10 :	Mike	(M)	20 :	Melissa	(F)

The letters "M","F", and "A" denote respectively a "Male adult" a "Female adult" and a "Juvenile".

The problem is the following: find a ranking of the chimpanzees from a pairwise domination tableau on the chimpanzees.

The pairwise comparison tableau C is defined by:

c_{ij} = 1 if the chimpanzee i dominates the chimpanzee j.

c_{ij} = 0 otherwise.

From the author's remarks on the social hierarchy on the internal dominations in the group, we have deduced the following C comparison tableau:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		0	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1
2	1		1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1
3	0	0		1	1	1	1	1	1	0	1	1	0	0	0	0	1	1	0	1
4	0	0	0		1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1
5	0	0	0	0		0	0	1	1	0	0	0	0	0	0	0	0	0	0	1
6	0	0	0	0	1		0	1	1	0	0	0	0	0	0	0	0	0	1	1
7	0	0	0	0	1	1		0	1	0	0	0	0	0	0	0	1	0	0	1
8	0	0	0	0	0	0	1		0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	1		0	0	0	0	0	0	0	0	0	0	0
10	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1
11	0	1	0	1	1	1	1	1	1	0		1	1	0	1	1	1	1	1	1
12	0	0	0	1	1	1	1	1	1	0	0		0	0	0	0	1	0	1	1
13	0	0	1	1	1	1	1	1	1	0	0	1		0	0	1	1	1	1	1
14	0	0	1	1	1	1	1	1	1	0	1	1	1		0	0	1	0	1	1
15	1	0	1	1	1	1	1	1	1	0	0	1	1	1		1	1	1	1	1
16	1	0	1	1	1	1	1	1	1	0	0	1	1	1	0		1	1	1	1
17	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0		0	0	1
18	0	0	0	0	1	1	1	1	1	0	0	1	0	1	0	0	1		1	1
19	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0	0	1	0		1
20	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	

This comparison tableau represents a binary relation which is
- Complete
-Non asymmetric: For instance the male number 13 dominates the male number 16 but reciprocally the male 16 dominates the male 13, there is a tie.
-Non transitive: For example the male number 7 dominates the juvenile number 9 who dominates the female number 8, but the male 7 does not dominate the female 8.

Thus the aggregation of this tableau by a complete order is not straightforward and the aggregation model is necessary. The ranking we are looking for is the optimal solution of the maximization problem:

$$\text{Max} \sum_i \sum_j c'_{ij} \; Y_{ij}$$

Y being a complete order

The maximum of this objective function is 173.
This maximum is obtained with the following optimal solution

1st - number 10 : Mike (M)		11th- number 12 : Huxley (M)	
2nd - number 11 : J.B. (M)		12th- number 4 : Flo (F)	
3th - number 2 : Goliath (M)		13th- number 19 : Faben (A)	
4th - number 15 : Rodolph (M)		14th- number 7 : William (M)	
5th - number 16 : Humphrey (M)		15th- number 17 : Evered (A)	
6th - number 1 : David (M)		16th- number 6 : Figan (A)	
7th - number 14 : Hugh (M)		17th- number 5 : Fifi (F)	
8th - number 13 : Leakey (M)		18th- number 9 : Flint (A)	
9th - number 3 : Mac Gregor (M)		19th- number 20 : Melissa (F)	
10th- number 18 : Worzle (M)		20th- number 8 : Olly (F)	

As we have seen in 2. this problem is equivalent to the minimization of the distance between Y and the C^k tableaux. Since there is only one individual tableau, this is equivalent to the minimization of the number of violations of the pairwise majority rule.

A violation of the pairwise majority rule occurs if:

$c'_{ij} < 0$ and $y_{ij} = 1$ or $c'_{ij} > 0$ and $y_{ij} = 0$

If the C' tableau is reordered according to the optimal ranking, the violations correspond to the positions with -1 in the upper triangular part of the tableau and to the positions with +1 in the lower triangular part of the tableau.
These positions have been circled for visualization.

	10	11	2	15	16	1	14	13	3	18	12	4	19	7	17	6	5	9	20	8
10	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	-1	0	1	1	1	(-1)	(-1)	1	(-1)	1	1	1	1	1	1	1	1	1	1	1
2	-1	-1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	-1	-1	-1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	-1	-1	-1	-1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	-1	(1)	-1	-1	-1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	-1	(1)	-1	-1	-1	-1	0	1	1	(-1)	1	1	1	1	1	1	1	1	1	1
13	-1	-1	-1	-1	(1)	-1	-1	0	1	1	1	1	1	1	1	1	1	1	1	1
3	-1	(1)	-1	-1	-1	-1	-1	-1	0	1	1	1	(-1)	1	1	1	1	1	1	1
18	-1	-1	-1	-1	-1	-1	(1)	-1	-1	0	1	(-1)	1	1	1	1	1	1	1	1
12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1	1	1	1	1	1	1	1	1
4	-1	-1	-1	-1	-1	-1	-1	-1	-1	(1)	-1	0	1	1	1	1	1	1	1	1
19	-1	-1	-1	-1	-1	-1	-1	-1	(1)	-1	-1	-1	0	1	1	(-1)	1	1	1	1
7	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1	1	1	1	1	(-1)
17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1	1	1	1	1
6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	(1)	-1	-1	0	1	1	1	1
5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1	1	1
9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1	1
20	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1
8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	(1)	-1	-1	-1	-1	-1	0

6. CONCLUSION

We have represented by optimization models and solved by optimization techniques many other aggregation problems. For instance:

- The individual and collective opinions are the same as above but the aggregation rule is changed. $c^!_{ij}$ is replaced by $sgn(c^!_{ij})$, that is to say if $c^!_{ij}$ is strictly positive it is replaced by +1 and replaced by -1 if it is strictly negative. The corresponding optimization model is the same as above: a binary linear program.

- The case of preference aggregation where the individual and collective opinions are rankings and various aggregation rules are considered. The corresponding optimization model is this time a continuous linear assignment program.

- In the case of aggregation of scales or notes, various aggregation rules are considered. The corresponding optimization model is this time a continuous non linear program with or without explicit solutions according to the aggregation rules.

All these models are detailed in [6].

7. REFERENCES

[1] D.BLACK,"The Theory of Comittees and Elections", Cambridge University Press, Cambridge (1958).

[2] V.J.BOWMAN, "Permutation polyhedra", Siam Journal of Applied Mathematics, (1975), Vol 22, No 4, 580

[3] J.de CANI, "Maximum likelihood paired comparison ranking by linear programming", Biometrika, (1969), Vol 56, No 3, 537.

[4] J.de CANI, " A branch and bound algorithm for maximum likelihood paired comparison ranking", Biometrika, (1972), Vol 59, No 1,131-135.

[5] J.KEMENY and L.SNELL, "Mathematical Models in the Social Sciences",The M.I.T Press,Cambridge USA,(1962)

[6] F.MARCOTORCHINO and P.MICHAUD,"Optimisation en Analyse Ordinale des Donnees", Masson, Paris, (1979).

[7] F.MARCOTORCHINO and P.MICHAUD,"Agregation de similarites en Classification Automatique",(to Appear: Revue de Statistiques Appliquees,Dunod, Paris (1981)).

[8] F.MARCOTORCHINO and P.MICHAUD, "Optimization in exploratory data analysis", in the proceedings of

the 5th Symposium uber Operations Research, Koln, (1980), (to appear by Physika Verlag ,1981).

[9] P.MICHAUD and F.MARCOTORCHINO, "Optimization in Ordinal Data Analysis", IBM-France Scientific Center Technical Report 001, (1978). (An important Part of this Paper has been presented at the International Symposium on Discrete Optimization, Vancouver August 1977).

[10] P.MICHAUD and F.MARCOTORCHINO,"Optimisation en Analyse des Donnees Relationnelles",in the Proceedings of the International Symposium on Data Analysis, sponsored by IRIA in Versailles, October 1979, North Holland,(1980), 655-670.

[11] P.MICHAUD and F.MARCOTORCHINO,"Modeles d'Optimisation en Analyse des Donnees Relationnelles", in Mathematiques et Sciences Humaines,Dunod Paris, (1979), No 67, 7-38.

[12] MPSX/370,"IBM Mathematical Programming System Extended 370", General Information Manual, GH19-1090.

[13] S.REGNIER, "Sur Quelques Aspects Mathematiques des Problemes de Classification Automatique", I.C.C Bulletin, Rome, (1965).